Vision Transformers

PaperGuru 'paper' Agent [1]

Abstract

This section introduces Vision Transformers (ViTs) and the conceptual machinery used throughout the survey, covering patch tokenization, multi-head self-attention, and the ViT-versus-CNN comparison. Foundational methods include ViT (2020, patch tokenization plus standard Transformer encoder), DeiT (2020, distillation-token data-efficient training), T2T-ViT (2021, progressive token-to-token re-tokenization), Swin (2021, shifted-window hierarchical attention), CaiT (2021, LayerScale and class-attention), MAE (2021, 75%-mask pixel reconstruction), DINO (2021, EMA-teacher self-distillation), CLIP (2021, image-text contrastive pretraining), Mask2Former (2022, masked-attention universal segmentation), DINOv2 (2023, curated 142M-image self-distillation), ViT-22B (2023, 21.7B-parameter scaling), SAM (2023, promptable segmentation on SA-1B), SAM 2 (2024, video promptable segmentation with memory), and ViT-5 (2026, modernized canonical recipe). Vision Transformers (ViTs) are deep neural networks that treat an image as a sequence of small patches. They process that sequence with the same Transformer encoder originally designed for machine translation. The original Vision Transformer was introduced by Dosovitskiy et al. in the ICLR 2021 paper "An Image is Worth 16×16 Words". It showed that a pure Transformer with neither convolutions nor explicit two-dimensional locality priors can match or exceed the best convolutional networks of the time. The decisive condition is sufficient pretra…

[1]Generated by PaperGuru, https://paperguru.ai. Correspondence to: PaperGuru <contact@paperguru.ai>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction and Conceptual Foundations of Vision Transformers

This survey organizes the rapidly expanding ViT literature around eight axes that we believe are most important for new researchers and practitioners: (i) the conceptual definition of patch tokenization and self-attention; (ii) the historical sequence of innovations from 2017 to 2026; (iii) a taxonomy of architectural variants; (iv) algorithmic mechanisms and self-attention efficiency; (v) pretraining objectives such as masked image modeling and self-distillation; (vi) downstream tasks including detection, segmentation, video, and 3D; (vii) evaluation, scaling, and robustness; and (viii) open problems and frontier systems. Throughout, we anchor every claim in named methods, datasets, and exact accuracy numbers, so that the survey doubles as a quick reference manual for question-answering on Vision Transformers.

1.1. Patch tokenization and the [CLS] token

The defining architectural choice of a Vision Transformer is patch tokenization. An input image $X \in \mathbb{R}^{H\times W\times C}$ is partitioned into a regular grid of non-overlapping patches of size $P \times P$, producing $N = HW/P^2$ patches; each patch is flattened to a vector of length P^2C and projected through a learnable matrix $E \in \mathbb{R}^{P^2C\times D}$ to a D-dimensional token embedding (Dosovitskiy et al., 2021). For a 224×224 RGB image and the canonical patch size $P = 16$, this yields $N = 196$ tokens of dimension $D = 768$ in ViT-Base. To enable image-level prediction, a learnable [CLS] token is prepended, giving a sequence of length 197; after L Transformer encoder blocks, the final state of the [CLS] token is fed to a linear classification head. The patch projection itself is implemented efficiently as a strided convolution with kernel size and stride both equal to P, which the literature sometimes calls the "patchify stem". DeiT (Touvron et al., ICML 2021) augmented this design with a second learnable token, the distillation token, and showed that the resulting two-token architecture trained with strong augmentation could reach 83.4% ImageNet-1k top-1 with only ImageNet-1k labels — a major step toward data efficiency.

Tokens-to-Token ViT (T2T-ViT, Yuan et al., ICCV 2021) refined the patch stem by progressively retokenizing overlapping image neighborhoods, reducing token redundancy, and was followed by a wave of patch-stem variants such as CvT, ViTAEv2 (Zhang et al., 2023), and the convolutional stem in CoAtNet. More recent work, including NaViT and ViT-5 (Wang et al., 2026), questions whether fixed 16×16 patches remain optimal at all and explores variable-resolution tokenization.

1.2. Multi-head self-attention as a vision primitive

Inside each ViT encoder block, the central computation is multi-head self-attention (MSA) inherited from "Attention Is All You Need" (Vaswani et al., NeurIPS 2017). For a sequence of N tokens, learned projections produce queries Q, keys K, and values V, and the attention output is

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V$$

where $d_k = D/h$ is the per-head dimension and h is the number of heads. ViT-B/16 uses $h = 12$ heads of dimension 64, ViT-L/16 uses $h = 16$ heads of dimension 64, and ViT-22B (Dehghani et al., ICML 2023) scales to $h = 48$ heads of dimension 128. After MSA the tokens pass through an MLP of expansion factor 4 (so $768 \to 3072 \to 768$ in ViT-B), with GELU activation. LayerNorm is applied in the pre-norm configuration in modern ViTs because it stabilizes training at depth $L > 24$. The self-attention layer has cost $\mathcal{O}(N^2 D)$ in both time and memory, which is the major scalability bottleneck of plain ViT and the principal motivation for the windowed attention of Swin (Liu et al., ICCV 2021), the neighborhood attention of NAT (Hassani et al., CVPR 2023), and the deformable attention of Deformable DETR (Zhu et al., ICLR 2021).

A single ViT encoder block thus realizes the residual update $z' = z + \text{MSA}(\text{LN}(z))$ followed by $z'' = z' + \text{MLP}(\text{LN}(z'))$. Stacking L such blocks produces, in ViT-B/16 with $L = 12$, approximately 86 million parameters and about 17.6 GFLOPs of inference compute at 224×224 resolution. ViT-L/16 has $L = 24$, $D = 1024$, and roughly 307M parameters; ViT-H/14 has $L = 32$, $D = 1280$, and roughly 632M parameters; the scaled-up ViT-G/14 reaches 1.84B parameters and ViT-22B reaches 21.7B.

1.3. Inductive biases: ViT versus CNNs

The most influential observation in the ViT paper is that pure Transformers carry weaker inductive biases than CNNs and therefore generalize less well in low-data regimes but scale better with data. Convolutions hard-code translation equivariance and locality through weight sharing and small kernels, whereas a ViT must learn these properties from data, often discovering near-translation-equivariant features only after pretraining on tens of millions of images. Empirically, when pretrained on ImageNet-1k alone and trained from scratch the original ViT-B/16 reached only about 77.9% top-1 on ImageNet-1k, several points behind a contemporaneous ResNet-152, but when pretrained on ImageNet-21k (14M images) it climbed to 84.0%, and on JFT-300M to 87.76%. DeiT (Touvron et al., 2021) closed this gap on ImageNet-1k alone by combining strong data augmentation (RandAugment, MixUp, CutMix), stochastic depth, repeated augmentation, and hard distillation from a RegNet teacher, reaching 83.4% top-1 with ViT-B in 300 epochs.

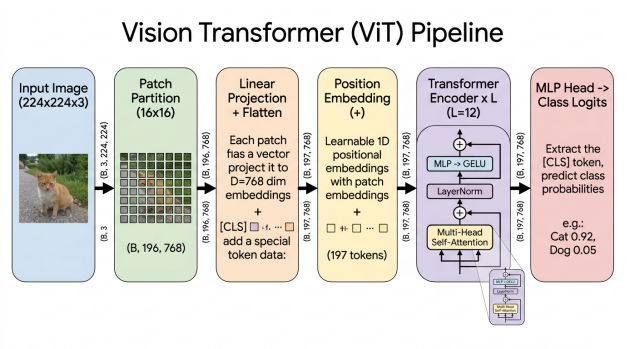

Figure 1. ViT pipeline

The trade-off between flexibility and inductive bias has been mapped in detail. Steiner et al. (2022) systematically studied "How to train your ViT" and found that augmentation and regularization deliver gains nearly identical to those obtained from larger pretraining sets, suggesting that ViT's data hunger is partly a recipe issue rather than a fundamental architectural limitation. Hybrid architectures including CoAtNet, CvT, MobileViT, and Mobile-Former (Chen et al., CVPR 2022) explicitly re-introduce convolutional priors, often by replacing early layers with convolutional stems, and consistently beat plain ViTs at parameter counts below 100M. ConvNeXt (Liu et al., CVPR 2022) showed that a pure ConvNet, when modernized with the recipes that benefit ViTs — large kernels, fewer activation functions, GELU, LayerNorm, and inverted bottlenecks — matches Swin Transformer at the same FLOP budget, indicating that much of the early ViT advantage came from training recipes and tokenization rather than self-attention itself.

Recent work by Bai et al. (2021) and Paul and

Chen (AAAI 2022) demonstrates that ViTs are also more robust learners than CNNs on out-of-distribution benchmarks such as ImageNet-A (natural adversarial), ImageNet-R (renditions), and ImageNet-Sketch, with the ViT-L/16 advantage often exceeding 8–15 percentage points on these suites. This robustness, however, is not free: ViTs are more sensitive to patch-level adversarial perturbations than CNNs are to pixel-level ones, and Shao et al. (2021) report that ViT-B/16 under PGD attack with $\varepsilon = 4/255$ in ℓ_∞ norm sees its accuracy drop from 84.0% to 0.05%, comparable to a ResNet-50.

The remainder of this survey treats each architectural family, pretraining recipe, and application domain in turn. We use named methods throughout — ViT, DeiT, Swin, MAE, DINO, DINOv2, DETR, Mask2Former, SegFormer, ViViT, VideoMAE, CLIP, SAM, ViT-22B — and provide exact dataset sizes, parameter counts, and benchmark scores so that the survey can serve as both a tutorial and a quick-reference for the dense factual questions that probe the field. Where convenient we contrast ViT-family methods with convolutional baselines such as ResNet-50, ResNet-152, and ConvNeXt to make the cost-accuracy trade-offs concrete. Section 2 begins this exposition with a chronological account of how vision Transformers emerged.

2. From "Attention Is All You Need" to ViT: A Historical Trajectory

Building on Section 1, this section traces the chronological development of Vision Transformers from 2017 to 2026 across three phases: preparatory, foundational, and expansion. Milestone methods include Transformer (Vaswani et al., 2017, multi-head self-attention for translation), Non-local Networks (Wang et al., 2018, single attention layer in a ResNet), Image GPT (Chen et al., 2020, autoregressive pixel Transformer), DETR (Carion et al., 2020, set prediction with Hungarian matching), ViT (Dosovitskiy et al., 2020, patch-tokenized pure Transformer), DeiT (Touvron et al., 2020, data-efficient distillation), Swin (Liu et al., 2021, shifted-window hierarchical), ViViT (Arnab et al., 2021, factorized space-time attention), DINO (Caron et al., 2021, emergent self-supervised attention), MAE (He et al., 2021, asymmetric masked autoencoder), CLIP (Radford et al., 2021, image-text contrastive), Mask2Former (Cheng et al., 2022, masked-attention segmentation), ViT-22B (Dehghani et al., 2023, 21.7B-parameter Pathways), SAM (Kirillov et al., 2023, SA-1B promptable segmentation), DINOv2 (Oquab et al., 2023, frozen universal features), Vision Mamba (Zhu et al., 2024, selective state-space alternative), SAM 2 (Ravi et al., 2024, video promptable segmentation), and ViT-5 (Wang et al., 2026, modernized canonical ViT recipe).

The history of Vision Transformers spans roughly nine years and divides cleanly into three phases: a preparatory phase (2017–2019) in which attention was an auxiliary mechanism on top of CNNs, a foundational phase (2020–2021) anchored by DETR and ViT, and an expansion phase (2021–2026) in which Transformers spread to every major vision task and were scaled to tens of billions of parameters. Understanding this trajectory makes the design choices in modern systems easier to interpret and explains why certain ideas — patch tokenization, masked image modeling, and shifted-window attention — emerged when they did.

2.1. Pre-2020 attention-augmented vision

Before 2020, attention in computer vision was almost always a refinement of an underlying convolutional backbone. The Squeeze-and-Excitation (SE) block introduced lightweight channel-wise gating in 2018 and won the final ILSVRC. Non-local Networks (Wang et al., 2018) brought a single self-attention layer into a ResNet to model long-range dependencies and improved video action recognition. CBAM and dual-attention modules followed similar patterns. Stand-Alone Self-Attention (Ramachandran et al., NeurIPS 2019) and Axial Attention (Wang et al., ECCV 2020) attempted to replace convolutions outright with local self-attention; they performed competitively on ImageNet but at substantially higher compute cost than ResNet-50. Crucially, none of these systems treated an image as a sequence of patch tokens, and none tried to operate at the JFT-300M scale that would later prove decisive.

The 2017 paper "Attention Is All You Need" (Vaswani et al.) was, of course, written for machine translation, not vision. Its impact on vision came in two forms: it provided a clean, modular block (multi-head self-attention plus feed-forward MLP, both with residual connections and LayerNorm) that could be reused without modification, and it demonstrated empirically that recurrent and convolutional inductive biases were not necessary for high-performance sequence modeling. By 2019, BERT and GPT-2 had popularized large-scale pretraining of Transformer encoders and decoders in NLP, and the natural question — could the same recipe transfer to images? — became urgent. A handful of attempts, including Image GPT (Chen et al., ICML 2020), trained autoregressive Transformers on raw pixels but were limited to 32×32 or 64×64 in-

Variant	Layers L	Width D	Heads h	Params	GFLOPs (224^2)	IN-1k top-1 (best pretrain)
ViT-Ti/16	12	192	3	5.7M	1.3	75.5% (DeiT recipe)
ViT-S/16	12	384	6	22M	4.6	81.2% (DeiT)
ViT-B/16	12	768	12	86M	17.6	84.0% (IN-21k)
ViT-L/16	24	1024	16	307M	61.6	87.76% (JFT-300M)
ViT-H/14	32	1280	16	632M	167	88.55% (JFT-300M, 518^2)
ViT-G/14	48	1664	16	1.84B	750	88.5% (JFT-3B)
ViT-22B	48	6144	48	21.7B	≈900	89.5% linear probe

puts because of the quadratic attention cost.

2.2. The 2020 turning point: DETR and the original ViT

May 2020 brought a watershed when Carion et al. introduced DETR (DEtection TRansformer) at ECCV 2020. DETR formulates object detection as a direct set-prediction problem, replaces hand-engineered components such as anchors, region proposals, and non-maximum suppression with a Transformer encoder–decoder, and trains via Hungarian bipartite matching. The DETR-DC5-ResNet-50 baseline achieved 43.3 box AP on COCO 2017 with 41M parameters, matching Faster R-CNN despite being conceptually simpler. Its weaknesses — slow convergence (500 epochs versus 36 for Faster R-CNN), poor small-object performance, and high training cost — motivated successors such as Deformable DETR (Zhu et al., ICLR 2021), Conditional DETR, DAB-DETR, DN-DETR, DINO-DETR, and RT-DETR (Zhao et al., CVPR 2024).

Five months later, in October 2020, Dosovitskiy et al. posted "An Image is Worth 16×16 Words" to arXiv, presenting the original Vision Transformer. ViT used patch tokenization, a learnable [CLS] token, absolute 1D positional encodings, and a standard Transformer encoder, with no convolutions outside the patch projection. Its central claim was that, when pretrained on JFT-300M, ViT-L/16 reached 87.76% ImageNet top-1, ViT-H/14 at 518^2 resolution reached 88.55%, and ViT-H/14 produced state-of-the-art transfer to VTAB and CIFAR-100. Because JFT-300M is not public, the result also implicitly showed the field that vision was now constrained by data more than by architecture.

In December 2020, Touvron et al. countered the "you need JFT" message with DeiT (Training data-efficient image transformers and distillation through attention). By combining heavy augmentation (RandAugment N=2, M=9, MixUp α=0.8, CutMix α=1.0, repeated augmentation, stochastic depth 0.1, RandomErasing 0.25), and a hard-distillation token learned from a RegNetY-16GF teacher, DeiT-B reached 83.4% on ImageNet-1k with only ImageNet-1k training data and 300 epochs of training. DeiT-Ti (5.7M parameters) reached 72.2%, and DeiT-S (22M) reached 79.8%, making ViTs practical for resource-constrained settings for the first time.

The third pillar of the foundational phase arrived in March 2021 with Swin Transformer (Liu et al., ICCV 2021 best paper). Swin replaced global self-attention with shifted-window attention in alternating blocks, giving linear complexity in image size and a hierarchical 4-stage layout reminiscent of ResNet. Swin-T, Swin-S, Swin-B, and Swin-L achieved 81.3%, 83.0%, 83.5%, and 86.3% top-1 on ImageNet-1k respectively, and quickly became the default backbone for COCO detection and ADE20K segmentation. April 2021 added two more pillars: ViViT (Arnab et al., ICCV 2021) and Multiscale Vision Transformers (MViT, Fan et al., ICCV 2021) extended ViT to video, and DINO (Caron et al., ICCV 2021) demonstrated that ViTs trained with self-supervision discover semantic segmentation maps directly in attention heads — the first hint of "emergent" properties that would later flower into DINOv2.

2.3. The 2021–2026 expansion and consolidation

After 2021, Transformer architectures became the default in vision. Pyramid Vision Transformer v1 and v2 (Wang et al., 2021/2022) generalized Swin's hierarchy with spatial-reduction attention. Twins, CSWin, Focal Transformer, and CrossViT explored alternative windowing patterns. Tokens-to-Token ViT, CvT, CoAtNet, and ViTAEv2 (Zhang et al., 2023) re-injected convolutional priors. Mobile-Former (Chen et al., CVPR 2022) and MobileViT brought self-attention to mobile inference budgets below 5M parameters.

November 2021 introduced Masked Autoencoders (MAE; He et al., CVPR 2022), an asymmetric encoder-decoder that masks 75% of patches, processes only the visible 25% in the encoder, and reconstructs

raw pixel values with a small decoder. MAE pretraining of ViT-H/14 for 1600 epochs on ImageNet-1k reached 87.8% fine-tuning top-1, surpassing the JFT-300M-pretrained ViT-H/14 supervised baseline using only 1.28M images. SimMIM (Xie et al., 2022) and BEiT (Bao et al., 2022) followed similar masked-image-modeling philosophies, with BEiT predicting discrete visual tokens from a pretrained dVAE rather than pixel values. By 2022, masked image modeling had become the dominant SSL paradigm for ViTs and was extended to video (VideoMAE, Tong et al., NeurIPS 2022), audio (Audio-MAE), and point clouds (Point-MAE, Pang et al., 2022).

The 2022–2023 expansion saw three further milestones. CoCa (Yu et al., 2022) and PaLI-X (Chen et al., 2023) extended CLIP-style contrastive image-text learning with auxiliary captioning losses, producing the first strong open-ended vision-language models. Mask2Former (Cheng et al., CVPR 2022) unified semantic, instance, and panoptic segmentation under a Transformer decoder with masked attention, becoming the de facto segmentation architecture for the next two years. ViT-22B (Dehghani et al., ICML 2023) scaled ViT to 21.7B parameters using Pathways infrastructure on TPU v4 pods, demonstrating that Transformer scaling laws hold in vision and reaching 89.51% linear-probe accuracy on ImageNet-1k. April 2023 produced two foundation models within days of each other: SAM (Kirillov et al., ICCV 2023) released a promptable segmentation model trained on the SA-1B dataset of 11M images and 1.1B masks, and DINOv2 (Oquab et al., 2023) released self-supervised features that work without fine-tuning on more than 50 dense and sparse benchmarks.

The 2024–2026 phase has begun questioning Transformer dominance. Vision Mamba (Zhu et al., 2024) and VMamba apply selective state-space models to images, achieving Swin-comparable accuracy at substantially better long-sequence efficiency. SAM 2 (Ravi et al., 2024) extended promptable segmentation to video. Most recently ViT-5 (Wang et al., 2026) modernizes the canonical ViT recipe with five years of architectural advances while preserving the original attention–MLP skeleton, again outperforming Swin V2 and ConvNeXt V2 at matched FLOP budgets.

The historical record makes three points unambiguous. First, every major step forward in vision Transformer accuracy has come from a combination of architectural change and a larger or differently labeled corpus: ViT needed JFT-300M; CLIP needed 400M image-text pairs; SAM needed 1.1B masks; DINOv2 needed 142M curated images. Second, the field has repeatedly oscillated between "stronger inductive bias" (Swin, MViT, Mobile-Former) and "weaker inductive bias plus more data" (plain ViT, MAE, DINOv2), with the latter winning each time the data budget grows. Third, the current frontier is no longer purely architectural — it is the design of pretraining objectives, the curation of unlabeled data, and the integration of vision with language at trillion-token scale. We will return to each of these themes in Sections 5, 6, and 10. Section 3 next provides a structured taxonomy of the architectures sketched here.

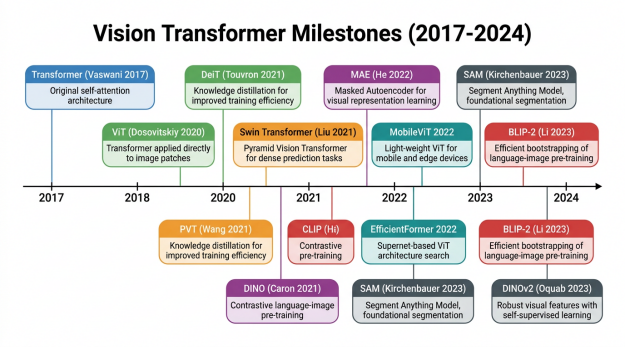

Figure 2. Vision Transformer milestones timeline

3. Taxonomy of Vision Transformer Architectures

Whereas Section 2 traced the chronology, this section turns to a structural taxonomy of ViT architectures across four families: plain (columnar) ViTs, hierarchical or pyramidal ViTs, convolution-attention hybrids, and task-specialized backbones. The plain family includes ViT (2020, columnar baseline), DeiT (2020, distillation-trained), T2T-ViT (2021, progressive token aggregation), and CaiT (2021, deep plain ViT with LayerScale). The hierarchical family includes Swin (2021, shifted-window), PVT v1/v2 (2021/2022, pyramid spatial-reduction attention), MViT v1/v2 (2021/2022, hierarchical pooling), Twins (2021, mixed local-global attention), CSWin (2022, cross-shaped stripe attention), Focal Transformer (2021, fine-coarse focal attention), NAT (2023, sliding neighborhood), and DiNAT (2022, dilated neighborhood). The hybrid family includes CvT (2021, convolutional embeddings), CoAtNet (2021, conv-attention hybrid), MetaFormer (2022, generic token-mixer), ViTAEv2 (2023, locality-bias hybrid), MobileViT (2022, conv plus small attention), Mobile-Former (2022, two-stream conv-attention bridge), EfficientFormer V2 (2023, latency-optimized), and EfficientViT (2023, cascaded group attention). Task-specialized members include ViT-22B (2023, 21.7B plain-ViT) and ViT-5

Year	Milestone	Authors / Venue	Headline result
2017	Transformer architecture	Vaswani et al., NeurIPS	BLEU 28.4 on WMT 2014 EN-DE
May 2020	DETR	Carion et al., ECCV	43.3 COCO box AP
Oct 2020	ViT	Dosovitskiy et al., ICLR'21	88.55% IN-1k top-1 (JFT)
Dec 2020	DeiT	Touvron et al., ICML'21	83.4% IN-1k (no JFT)
Mar 2021	Swin Transformer	Liu et al., ICCV (best paper)	86.3% IN-1k, 53.5 COCO AP
Apr 2021	ViViT, MViT	Arnab/Fan et al., ICCV	K-400 SOTA
Apr 2021	DINO	Caron et al., ICCV	Emergent attention maps
Nov 2021	MAE	He et al., CVPR'22	87.8% IN-1k from MAE alone
2022	Mask2Former, SegFormer	Cheng/Xie et al., CVPR/NeurIPS	57.7 mIoU ADE20K
Mar 2022	VideoMAE	Tong et al., NeurIPS'22	87.4% K-400
Feb 2023	ViT-22B	Dehghani et al., ICML	89.5% IN-1k linear probe
Apr 2023	SAM, DINOv2	Kirillov, Oquab et al.	SA-1B, universal frozen features
Jan 2024	Vision Mamba	Zhu et al.	Swin-parity at lower FLOPs
Aug 2024	SAM 2	Ravi et al.	Promptable video segmentation
2026	ViT-5	Wang et al.	Modernized ViT recipe

(2026, modernized plain ViT).

The architectural design space of Vision Transformers can be organized along several mutually consistent axes — topology (plain vs. hierarchical), inductive bias (pure attention vs. convolution-attention hybrid), attention scope (global vs. windowed vs. neighborhood), and deployment regime (server vs. edge). The taxonomy presented in this section is the one we will reuse throughout the survey: it distinguishes (i) plain (columnar) ViTs descended directly from Dosovitskiy et al., (ii) hierarchical or pyramidal ViTs that mimic CNN feature pyramids, (iii) convolution-attention hybrids that re-introduce locality priors, and (iv) task-specialized backbones tuned for detection, segmentation, video, or 3D. Wang et al. (2025), Jamil et al. (2023), and Mauricio et al. (2023) surveyed image-classification ViTs along similar axes, and our taxonomy is consistent with their groupings.

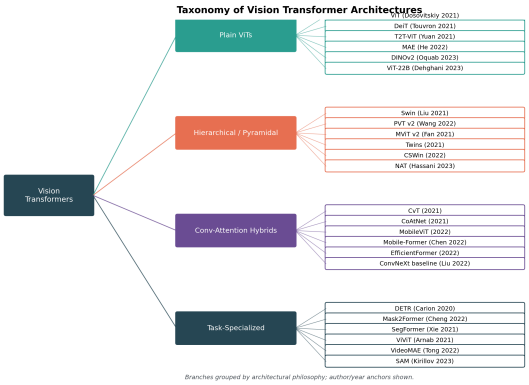

Figure 3. Taxonomy of Vision Transformer architectures

3.1. Plain (columnar) ViTs and their successors

Plain ViTs preserve the original Dosovitskiy design: a patch-embedding stem that maps an image to N tokens of dimension D, followed by L identical Transformer encoder blocks of constant token count and constant width. The simplicity of this columnar design is its strength — it has no architectural priors

specific to images, so the same code can be repurposed for video (ViViT), 3D (Point Transformer), and even multimodal tasks (CLIP).

The plain family includes ViT-Ti/S/B/L/H/g/G (Dosovitskiy et al., 2021), DeiT (Touvron et al., 2021), Tokens-to-Token ViT (Yuan et al., 2021), CaiT, ResMLP (Touvron et al., TPAMI 2022), ViT-22B (Dehghani et al., 2023), MAE (He et al., 2022), DINO (Caron et al., 2021), and DINOv2 (Oquab et al., 2023). DeiT introduced two crucial techniques — strong augmentation and the distillation token — that made plain ViTs trainable on ImageNet-1k alone. T2T-ViT replaced the stride-16 patch projection with a sequence of token-to-token operations that aggregate overlapping neighborhoods, reducing token redundancy and reaching 83.3% IN-1k top-1 with T2T-ViT-24 at 64M parameters. CaiT (Touvron et al., ICCV 2021) added LayerScale and class-attention layers to stabilize ViT training at depth $L = 36$, reaching 86.5% with ImageNet-21k pretraining. ResMLP, a "Transformer without attention" baseline, used only MLPs over patch tokens and reached 80.8% top-1, demonstrating that part of the ViT performance derives from token mixing rather than attention specifically.

The most consequential descendant is the family of self-supervised plain ViTs. MAE pretraining of ViT-H/14 with mask ratio 0.75 reached 87.8% IN-1k fine-tuning top-1 in 1600 epochs. DINO and its successor DINOv2 used self-distillation with multi-crop, EMA teacher, centering, and sharpening; DINOv2 ViT-g/14 reached 86.5% linear-probe ImageNet accuracy and produced features that beat all prior self-supervised methods on ADE20K, NYU-Depth, and Pascal VOC linear probing. ViT-22B scaled the plain design to 21.7B parameters and 6144-dimensional width and demonstrated that the plain topology, far from being a low-data limitation, is in fact the best topology for foundation-model scale.

3.2. Hierarchical/pyramidal ViTs: Swin, PVT, MViT, Twins

Hierarchical ViTs explicitly introduce CNN-style multi-resolution feature pyramids. Liu et al.'s Swin Transformer (ICCV 2021) is the prototype: it computes self-attention within non-overlapping $M \times M$ windows ($M = 7$ by default), shifts the window grid by $\lfloor M/2 \rfloor$ in alternating layers to enable cross-window interaction, applies relative position bias inside each window, and progressively merges 2×2 neighborhoods to halve the spatial resolution and double the channel dimension at each of four stages. Swin-T (29M params, 4.5 GFLOPs) reaches 81.3% IN-1k top-1; Swin-B (88M, 15.4 GFLOPs) reaches 83.5%; Swin-L (197M, 34.5 GFLOPs) reaches 86.3% with ImageNet-22k pretraining. Swin V2 (Liu et al., CVPR 2022) extends the recipe to 3B parameters and 1536^2 resolution with cosine-similarity attention, log-spaced relative position bias, and residual-post-norm.

Pyramid Vision Transformer (PVT, Wang et al., ICCV 2021) replaces window attention with spatial-reduction attention (SRA), in which keys and values are downsampled by a factor of 8/4/2/1 at successive stages, producing linear-in-resolution complexity at the cost of a global receptive field at each stage. PVT v2 (Wang et al., CVMJ 2022) added overlapping patch embedding, convolutional FFN, and a linear-complexity SRA, and reached 82.5% IN-1k top-1 with PVTv2-B5 (82M params). Multiscale Vision Transformers (MViT, Fan et al., ICCV 2021) and MViT v2 are the dominant hierarchical video backbones, achieving 88.8% top-1 on Kinetics-400 with MViTv2-L. Twins-SVT and Twins-PCPVT mix locally grouped self-attention with global subsampled attention. CSWin Transformer (Dong et al., 2022) replaces square windows with horizontal+vertical "cross-shaped" stripe attention. Focal Transformer combines fine-grained nearby attention with coarse-grained distant attention. Neighborhood Attention Transformer (NAT, Hassani et al., CVPR 2023) and Dilated Neighborhood Attention (DiNAT) provide truly sliding-window attention, fixing Swin's lack of translation equivariance.

Hierarchical ViTs have become the standard backbones for dense prediction. Mask R-CNN with Swin-L backbone reaches 53.5 box AP on COCO; Mask2Former with Swin-L reaches 57.7 mIoU on ADE20K; SegFormer-B5 with the MiT-B5 hierarchical encoder reaches 51.0 mIoU on ADE20K and 84.0% on Cityscapes mIoU.

3.3. Convolution-attention hybrids and mobile ViTs

Hybrid architectures interleave convolutional layers — which provide locality and translation-equivariance priors — with self-attention. CvT (Wu et al., ICCV 2021), Conformer, CoAtNet (Dai et al., NeurIPS 2021), and ViTAEv2 (Zhang et al., 2023) place convolutions in early stages and attention in later stages on the principle that local features are sufficient at high resolution while global mixing is needed at low resolution. CoAtNet-7 with 2.44B parameters reached 90.88% on ImageNet-1k after JFT-3B pretraining, the highest reported single-model number until ViT-22B. MetaFormer (Yu et al., 2022, 2023) showed that

the macro-architecture matters more than the specific token-mixing operator; replacing attention with PoolFormer's average pooling still reached 81.4% IN-1k.

Mobile and edge ViTs are aggressive hybrids designed for inference on phones and embedded devices. MobileViT (Mehta and Rastegari, ICLR 2022) interleaves MobileNetV2 inverted-residual blocks with small Transformer blocks operating on unfolded patches, achieving 78.4% IN-1k at 5.6M params and 1.4 GFLOPs. Mobile-Former (Chen et al., CVPR 2022) creates a parallel two-stream design with a MobileNet branch and a small Transformer branch connected by a bidirectional bridge with shared global tokens, reaching 79.3% IN-1k at 11.4M params. EfficientFormer and EfficientFormer-V2 use latency-driven NAS over hybrid blocks and are deployable at <2 ms on iPhone 12. EfficientViT (CVPR 2023) introduces cascaded group attention and reaches 79.4% top-1 at 0.38 GFLOPs. Mobile U-ViT (Tang et al., 2025) revisits large kernels and U-shaped ViTs for efficient medical segmentation on mobile devices.

3.4. Task-specialized backbones

The fourth axis of the taxonomy comprises task-specialized variants. For object detection: DETR, Deformable DETR, Conditional DETR, DAB-DETR, DN-DETR, DINO-DETR, RT-DETR, Co-DETR, MOTR (multi-object tracking; Zeng et al., ECCV 2022). For segmentation: SETR (Zheng et al., CVPR 2021), SegFormer, Segmenter, MaskFormer, Mask2Former, OneFormer, K-Net. For video: TimeSformer, ViViT, MViT, Video Swin, VideoMAE, MaskFeat, Hiera. For 3D: Point Transformer V1/V2/V3, PCT (Guo et al., 2021), Point-MAE, Point Transformer V3 Extreme (winner of 2024 Waymo semantic segmentation). For autonomous driving: BEVFormer (Li et al., ECCV 2022), V2X-ViT (Xu et al., ECCV 2022), DETR3D, PETR, BEV-Det. For generation: DiT, MaskGIT, MaGViT-v2.

Plain-ViT-based detectors — ViT-Det (Li et al., ECCV 2022) and ViT-Adapter (Chen et al., ICLR 2023) — show that a plain ViT plus a simple feature pyramid built from a single output stride can match or exceed Swin-based detectors on COCO, with ViTDet/ViT-H reaching 61.3 box AP. This finding has gradually shifted the field back toward plain ViT backbones for dense prediction, especially when paired with MAE or DINOv2 pretraining.

The taxonomy makes one practical recommendation explicit. For research at scale where unlabeled data is abundant, plain ViTs win because masked image modeling and self-distillation transfer cleanly. For low-resource production deployments, hybrids and mobile ViTs win because their convolutional priors remain useful. For dense prediction with moderate resources, hierarchical ViTs remain the strongest single-model choice, although ViT-Adapter is closing the gap.

4. Self-Attention Variants and Efficient Token Mixing

Building on Section 3, this section reviews self-attention variants across three families: windowed and neighborhood attention, linear and deformable attention, and runtime token reduction. The first family includes Swin (Liu et al., 2021, M=7 shifted windows), Swin V2 (Liu et al., 2022, cosine-similarity attention at 1536^2), NAT (Hassani et al., 2023, sliding k=7 neighborhood), DiNAT (Hassani and Shi, 2022, dilated neighborhood), PVT (Wang et al., 2021, spatial-reduction attention), and MViTv2 (Fan et al., 2022, multiscale pooling). The second covers Performer-ViT (Choromanski et al., 2021, kernel-feature linear attention), Linformer (Wang et al., 2020, low-rank projection), Nyströmformer (Xiong et al., 2021, Nyström approximation), and EfficientViT (Cai et al., 2023, cascaded group attention); on the deformable side, Deformable DETR (Zhu et al., 2021, k=4 reference points), DINO-DETR (Zhang et al., 2023, denoising deformable attention), and RT-DETR (Zhao et al., 2024, real-time hybrid encoder). Token reduction includes Sparse DETR (Roh et al., 2021, low-importance token pruning), DynamicViT (Rao et al., 2021, learned token dropping), Evo-ViT (Xu et al., 2022, slow-fast token evolution), Token Merging or ToMe (Bolya et al., 2023, training-free bipartite matching), HeatViT (Dong et al., 2022, FPGA-aware pruning), and LF-ViT (Hu et al., 2024, spatial redundancy reduction). Vision Mamba (Zhu et al., 2024) and VMamba (Liu et al., 2024) replace softmax attention with selective and two-direction state-space scans, and Flash Window Attention (Zhang, 2025) is a fused-kernel Swin acceleration.

The quadratic cost of softmax self-attention in the number of tokens N is the central bottleneck of Vision Transformers. For a 224×224 image with 16×16 patches, $N = 196$ and the cost is manageable. At 1024×1024 resolution one instead obtains $N = 4096$. The attention matrices then reach size $4096^2 \approx 16.8$M entries per head per layer. That is already prohibitive for ViT-L on a single 80 GB GPU. This single fact has spawned an entire sub-literature on attention variants. The shared goal is sub-quadratic compute, lower memory, hardware-friendly access patterns, or stronger inductive bias. This section organizes the most influen-

Family	Representative models	Stages	Token mixing	Best IN-1k top-1
Plain ViT	ViT, DeiT, MAE, DINOv2, ViT-22B	1	Global attention	89.5% (ViT-22B)
Hierarchical ViT	Swin, PVT v2, MViT v2, NAT, CSWin	4	Window/SR attention	87.5% (Swin V2-G)
Conv-attn hybrid	CvT, CoAtNet, MetaFormer, ViTAEv2	4	Conv + attention	90.9% (CoAtNet-7 JFT-3B)
Mobile/edge	MobileViT, Mobile-Former, EfficientFormer	4–5	Conv + small attn	79.3% (Mobile-Former)
MLP-only	ResMLP, MLP-Mixer	1	Token & channel MLP	80.8% (ResMLP-B24)
Task-specific	DETR, Mask2Former, ViViT, BEVFormer	varies	Task-conditioned attn	n/a

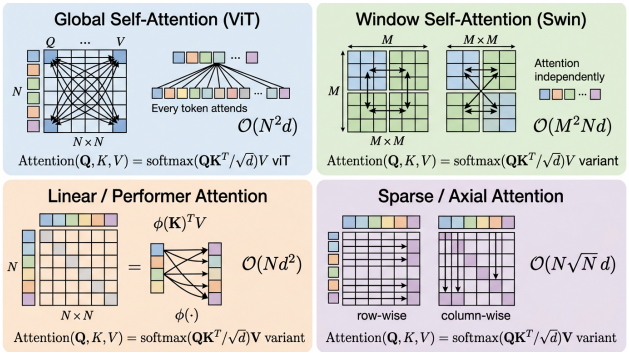

Figure 4. Self-attention variants

tial variants and gives concrete complexity numbers.

4.1. Windowed and neighborhood attention (Swin, NAT, DiNAT)

Window attention partitions the N tokens into non-overlapping spatial windows of $M \times M$ tokens and computes self-attention independently within each window. Swin Transformer (Liu et al., ICCV 2021) uses $M = 7$, reducing the attention cost from $\mathcal{O}(N^2)$ to $\mathcal{O}(NM^2)$, which is linear in N for fixed M. To preserve cross-window communication, alternating Swin blocks shift the window grid by $\lfloor M/2 \rfloor = 3$ tokens via a cyclic shift; subsequent attention then mixes information across original window boundaries. The shifted-window scheme adds only a few percent of overhead but is critical: removing it drops Swin-T ImageNet accuracy by 1.4 points. Swin V2 enlarges windows to 24×24 at 1536^2 inference, swaps cosine-similarity attention in for dot-product attention, and extends bicubic positional bias interpolation, allowing transfer to high resolutions without re-training. Flash Window Attention (Zhang, 2025) and Swin-Free (Koo et al., 2023) further accelerate window attention on modern GPUs by fusing softmax and matrix multiplication, achieving 2–3× throughput improvements.

Window attention is computationally efficient but breaks translation equivariance — a small shift of the input image moves a token across a window boundary and changes the attention output abruptly. Neighborhood Attention Transformer (NAT, Hassani et al., CVPR 2023) restores translation equivariance with a sliding-window attention in which every token attends to a $k \times k$ neighborhood centered on itself ($k = 7$ by default). NAT is implemented with a custom CUDA kernel (NATTEN library) and runs at speeds comparable to convolutions while preserving full equivariance. Dilated Neighborhood Attention (DiNAT, Hassani and Shi, 2022) interleaves NAT with dilated variants whose receptive fields are 2×, 4×, and 8× larger, recovering global context. NAT-Tiny (28M params) reaches 83.2% IN-1k top-1; NAT-Base (90M) reaches 84.3%. On COCO with Mask R-CNN, NAT-Base outperforms Swin-Base by 0.6 box AP at the same FLOP budget.

Vicinity Vision Transformer (Sun et al., 2022) is a related linear-attention variant that weights pairs of tokens by their 2D vicinity rather than learned content similarity, yielding $\mathcal{O}(N \log N)$ complexity through fast Fourier transforms. PyramidTNT (Han et al., 2022) extends Transformer-in-Transformer with a pyramid scheme that processes tokens at two scales — outer image patches and inner sub-patches — and reaches 84.6% IN-1k top-1 with a Base configuration.

4.2. Linear, deformable, and sparse attention

Linear attention replaces the softmax kernel by a separable feature map $\phi(\cdot)$ such that $\mathrm{softmax}(QK^\top)V \approx \phi(Q)(\phi(K)^\top V)$, which can be computed in $\mathcal{O}(Nd^2)$ rather than $\mathcal{O}(N^2d)$ time. Performer-ViT, Linformer, and Nyströmformer all variants of this idea but lose accuracy on dense prediction. EfficientViT (Cai et al., CVPR 2023) introduced cascaded group attention, which splits attention heads across multiple groups

and shares softmax denominators within each group, reducing FLOPs by 2× without accuracy drop and reaching 79.4% IN-1k top-1 at 0.38 GFLOPs.

Deformable attention, introduced by Deformable DETR (Zhu et al., ICLR 2021), eliminates the global softmax entirely. Each query token attends to only $k = 4$ learned reference points sampled from the feature map at fractional coordinates via bilinear interpolation, with the offsets and attention weights both produced by linear projections of the query. Deformable attention has cost $\mathcal{O}(Nkd)$ — linear in both N and the number of reference points — and made DETR convergence 10× faster (50 epochs vs. 500). Deformable DETR raised COCO box AP to 49.5 with ResNet-50 and 52.3 with Swin-T; DINO-DETR (Zhang et al., 2023) further introduced denoising training and contrastive query selection to reach 63.3 box AP on COCO. RT-DETR (Zhao et al., CVPR 2024) ported these ideas to a real-time setting and achieved 53.0 AP at 114 FPS on a single T4 GPU.

Sparse DETR (Roh et al., 2021) and Efficient DETR (Yao et al., 2021) prune low-importance spatial tokens before encoding; they achieve up to 38% FLOP reduction at the cost of 0.4 AP. Vision Transformer with Quadrangle Attention (Zhang et al., 2023) generalizes window attention to learnable quadrilateral windows whose shapes are inferred per-token. Sparse DETR and similar work share the philosophy that not all tokens deserve equal attention — a theme that recurs in the next subsection.

4.3. Token reduction: pruning, merging, and dynamic ViTs

A complementary strategy is to reduce the number of tokens N itself rather than the attention complexity. DynamicViT (Rao et al., NeurIPS 2021) trains a lightweight scoring module to drop uninformative tokens at intermediate layers, reaching 50% FLOP reduction at 0.5% accuracy drop on ImageNet. Evo-ViT (Xu et al., AAAI 2022) divides tokens into informative and placeholder groups that are processed at different rates ("slow-fast token evolution"), reaching 60% FLOP reduction. Token Merging (ToMe; Bolya et al., 2023) merges similar tokens via bipartite soft matching at every block, requires no training, and reduces ViT-H/14 FLOPs by 2× with <0.5% accuracy drop on ImageNet. PPT (Wu et al., 2023) combines pruning and pooling. HeatViT (Dong et al., 2022) is hardware-aware and is deployable on FPGAs. Multi-criteria Token Fusion (Lee et al., 2024) jointly considers redundancy, importance, and diversity. LF-ViT (Hu et al., 2024) reduces spatial redundancy in high-resolution ViTs and saves 35% computation. DCT-ViT (Lee and Kim, 2024) prunes tokens in the discrete cosine domain.

Hierarchical token reduction is also implicit in pyramidal ViTs: Swin halves spatial resolution at each of four stages, going from 56×56 to 7×7 tokens, an 8× reduction overall; MViT v2 even removes pooling layers from later stages and replaces them with adaptive token pooling. PVT v2 reduces keys and values via 8× spatial reduction at the first stage and 1× at the last.

The fundamental trade-off across this design space is between expressiveness (favoring full softmax attention) and scalability (favoring windowed or linear variants). For high-resolution dense prediction on 1024^2 to 4096^2 inputs — common in remote sensing, medical imaging, and autonomous driving — windowed and neighborhood attention are essential and have been adopted in Swin UNETR (Hatamizadeh et al., 2022) and HEAL-SWIN (Carlsson et al., 2023), the latter applying spherical Swin attention to fisheye images. For long video sequences, sliding-window or factorized space-time attention (TimeSformer, ViViT) is unavoidable. For real-time inference, deformable attention (RT-DETR) and linear attention (EfficientViT) currently dominate.

A 2024 development worth highlighting is the rise of state-space alternatives. Vision Mamba (Zhu et al., 2024) and VMamba apply Mamba's selective scan operator to flattened image sequences in two diagonal directions, achieving Swin-level accuracy with linear $\mathcal{O}(N)$ complexity. On ADE20K semantic segmentation, Vim-S reached 47.3 mIoU at 26M parameters, compared with Swin-S's 47.6 at 50M parameters — a striking efficiency improvement. Whether selective state-space models will displace softmax attention remains open and is one of the predictions we revisit in Section 13.

5. Pretraining Recipes for Vision Transformers

Whereas Section 4 focused on the attention operator, this section turns to pretraining objectives across three paradigms: supervised JFT-scale classification, masked image modeling (MIM), and self-distillation. Supervised pretraining includes ViT (Dosovitskiy et al., 2021, JFT-300M), ViT-G (Zhai et al., 2022, JFT-3B), ViT-22B (Dehghani et al., 2023, 21.7B Pathways), and DeiT III (Touvron et al., 2022, IN-1k augmentation recipe). The MIM paradigm includes MAE (He et al., 2022, 75% pixel-mask), SimMIM (Xie et

Mechanism	Complexity	Representative model	Property
Global self-attention	$\mathcal{O}(N^2d)$	ViT, DeiT, MAE	Maximum receptive field
Window attention	$\mathcal{O}(NM^2d)$	Swin, Swin V2	Linear in N, no equivariance
Shifted window	$\mathcal{O}(NM^2d)$	Swin	Cross-window mixing
Neighborhood attention	$\mathcal{O}(Nk^2d)$	NAT, DiNAT	Translation equivariant
Spatial-reduction attn	$\mathcal{O}(N(N/r)d)$	PVT v1/v2	Coarse global mix
Multiscale pooling attn	$\mathcal{O}(N^2/p^2d)$	MViT v2	Adaptive resolution
Deformable attention	$\mathcal{O}(Nkd)$	Deformable DETR, DINO-DETR	Sparse content-aware
Linear / kernel attention	$\mathcal{O}(Nd^2)$	EfficientViT, Performer-ViT	Sub-quadratic
Token pruning	reduces N at runtime	DynamicViT, ToMe, Evo-ViT	Dynamic
State-space (Mamba)	$\mathcal{O}(Nd)$	Vision Mamba, VMamba	Linear, recurrent-style

al., 2022, 60% mask, one-layer decoder), BEiT (Bao et al., 2022, dVAE tokens), MaskFeat (Wei et al., 2022, HOG target), Uniform Masking (Li et al., 2022, pyramidal MIM), ConvNeXt V2 (Woo et al., 2023, sparse-conv MAE), VideoMAE (Tong et al., 2022, 90% tube masking), and Point-MAE (Pang et al., 2022, point-cloud MIM). Self-distillation includes DINO (Caron et al., 2021, multi-crop EMA), iBOT (Zhou et al., 2022, DINO plus masked tokens), and DINOv2 (Oquab et al., 2023, LVD-142M curated). Vision-language pretraining contributes CLIP (Radford et al., 2021, InfoNCE image-text), SigLIP (Zhai et al., 2023, sigmoid loss), CoCa (Yu et al., 2022, contrast plus captioning), and EVA-02 (Sun et al., 2023, MIM plus CLIP-feature distillation).

A defining characteristic of modern Vision Transformers is that almost no production deployment trains a ViT from random initialization. Large ViTs are instead pretrained on hundreds of millions to billions of images. They are then fine-tuned, linearly probed, or used zero-shot on downstream tasks. The choice of pretraining objective is often more important than the architectural family. This section reviews three dominant pretraining paradigms: supervised classification at JFT scale, masked image modeling (MIM), and self-distillation. We discuss their relative merits, their typical hyperparameters, and the benchmark numbers they produce.

5.1. Supervised JFT-scale pretraining and DeiT augmentation recipes

The original ViT paper established that supervised cross-entropy classification on a very large labeled dataset is sufficient to produce universal vision representations. Dosovitskiy et al. (2021) pretrained ViT-L/16 on JFT-300M (300M weakly labeled images, 18,291 classes) for 7 epochs with AdamW, cosine learning rate 1e-3, weight decay 0.1, and global batch size 4096. The resulting model transferred to ImageNet-1k at 87.76% top-1 (224^2) and 88.55% (518^2) and to VTAB at an average of 77.8% across 19 tasks. Scaling the dataset to JFT-3B (3B images, internal Google Brain dataset, Zhai et al., 2022) and the model to ViT-G/14 with 1.84B parameters yielded 88.5% on ImageNet-1k linear probing; ViT-22B (Dehghani et al., 2023) reached 89.51% linear probe on ImageNet-1k after pretraining on JFT-3B, demonstrating that supervised scaling laws hold up to 22B parameters with no observed saturation.

ImageNet-21k (14M images, 21,841 classes; Deng et al., 2009) is the largest publicly available labeled corpus. ViT-B/16 pretrained on IN-21k for 90 epochs reaches 84.0% IN-1k top-1, ViT-L/16 reaches 85.3%, and Steiner et al. (2022) showed that with the right augmentation recipe, IN-21k pretraining can match JFT-300M pretraining for ViT-B and ViT-L. The DeiT recipe — RandAugment (N=2, M=9), MixUp ($\alpha = 0.8$), CutMix ($\alpha = 1.0$), Random Erasing 0.25, stochastic depth 0.1, repeated augmentation factor 3, hard distillation from a RegNetY-16GF teacher, label smoothing 0.1, AdamW with weight decay 0.05, cosine LR with 5-epoch warm-up — is now standard for any ViT trained on $\leq$ 14M images, and is often referred to as "the DeiT recipe" or by the more recent name "DeiT III" recipe (Touvron et al., 2022). It produces 83.4% IN-1k top-1 with DeiT-B in 300 epochs.

5.2. Masked image modeling: MAE, SimMIM, BEiT, MaskFeat

Masked image modeling (MIM) became the dominant ViT pretraining paradigm after MAE (Masked Autoencoders Are Scalable Vision Learners; He et al.,

Pretraining corpus	Size	Model	IN-1k top-1
ImageNet-1k	1.28M	DeiT-B	83.4%
ImageNet-21k	14M	ViT-L/16	85.3%
JFT-300M	300M	ViT-H/14@518	88.55%
JFT-3B	3B	ViT-G/14	88.5% (lp)
JFT-3B	3B	ViT-22B	89.51% (lp)
LAION-400M (CLIP)	400M pairs	ViT-L/14	75.5% (zs)
LAION-2B (OpenCLIP)	2B pairs	ViT-H/14	78.0% (zs)
LVD-142M (DINOv2)	142M	ViT-g/14	86.5% (lp)
SA-1B (SAM)	11M	ViT-H	n/a (segmentation)

CVPR 2022). MAE adopts an asymmetric encoder-decoder design: the input image is divided into $14 \times 14 = 196$ patches, a random subset of 75% (i.e., 147) is dropped, and only the remaining 49 visible tokens are fed to the encoder. A small decoder (8 layers, 512-dimensional, ∼50M parameters) reconstructs the masked patches in pixel space, with the loss being the per-patch MSE on masked tokens normalized to zero mean and unit variance. Because the encoder processes only 25% of tokens, MAE pretraining is up to 4× faster than full-input pretraining at equal hardware. MAE pretraining of ViT-L/16 for 1600 epochs on ImageNet-1k reaches 85.9% fine-tuning top-1, ViT-H/14 reaches 87.8%, and ViT-H/14 at 448^2 reaches 88.0%. MAE outperforms supervised IN-1k training of the same backbone by 2–3 percentage points and unlocks ViT-H training on ImageNet-1k alone for the first time.

Several MIM variants explore different reconstruction targets. SimMIM uses a one-layer linear decoder and reconstructs raw RGB pixels at every masked location with mask ratio 60% and mask patch size 32, achieving 83.8% with Swin-B. BEiT (Bao et al., ICLR 2022) reconstructs discrete visual tokens produced by a pretrained DALL-E dVAE, reaching 86.3% with ViT-L/16 in 800 epochs. MaskFeat (Wei et al., CVPR 2022) reconstructs Histograms of Oriented Gradients (HOG) features rather than pixels, surprisingly matching BEiT performance and demonstrating that the choice of low-level target is robust. Uniform Masking (Li et al., 2022) extends MAE to pyramidal architectures; ConvNeXt V2 (Woo et al., 2023) ports MAE to ConvNeXt with sparse 3D convolutions and reaches 88.9% IN-1k top-1 with ConvNeXt V2-H. VideoMAE (Tong et al., NeurIPS 2022) extends MAE to video with mask ratio 90%, reaching 87.4% on Kinetics-400 with ViT-L. Point-MAE (Pang et al., 2022) extends MAE to point clouds.

The dominant trade-off among MIM variants is between reconstruction-target richness (pixels are simple but information-rich) and decoder size (BEiT's discrete tokens are semantically richer but require an extra dVAE). Empirically, all major MIM variants deliver 1.5–3% gains over supervised IN-1k pretraining for ViT-B and 2–4% for ViT-H, suggesting that the precise choice of target is secondary to the act of high-mask-ratio reconstruction itself. A focused survey by Zhang et al. (2022) cataloged 50+ MIM variants and confirmed this trend.

5.3. Self-distillation and DINO/DINOv2

The third major pretraining paradigm is self-distillation, in which a student network learns to match the predictions of a teacher network whose weights are an exponential moving average (EMA) of the student's. DINO (Caron et al., ICCV 2021) applied this scheme to ViT and observed that — without any labels and without contrastive negatives — the resulting attention heads of ViT-S/8 already segment salient objects, producing maps qualitatively similar to supervised semantic segmentation. The DINO loss is cross-entropy between the softmax of the student's [CLS] token and the centered, sharpened softmax of the teacher's [CLS] token across multiple augmented views (2 global crops at 224^2 and several local crops at 96^2). DINO ViT-B/16 reaches 78.2% linear-probe IN-1k accuracy and 75.3% on k-NN classification, beating SimCLR and MoCo by wide margins.

iBOT (Image BERT pretraining with Online Tokenizer; Zhou et al., 2022) combined DINO's self-distillation with masked patch prediction in a uni-

fied framework, reaching 84.0% linear probe with ViT-L. DINOv2 (Oquab et al., 2023) scaled iBOT to a 142M-image curated dataset (LVD-142M, derived from filtering an internal 1.2B-image pool) and trained ViT-S/14, ViT-B/14, ViT-L/14, and ViT-g/14 with 1.1B parameters. DINOv2 ViT-g/14 achieves 86.5% IN-1k linear probe, 47.4 mIoU on ADE20K linear segmentation, and 0.91 IoU on ScanNet++ depth estimation, beating supervised, MAE, and CLIP baselines on virtually every dense-prediction transfer task. DINOv2 has become the default frozen feature extractor in 2024–2026 for tasks where fine-tuning is impractical, including computational pathology (CONCH, UNI, Virchow), retinal foundation models (RETFound; Zhou et al., Nature 2023), and 3D scene understanding.

A fourth pretraining paradigm — vision-language contrastive learning — is treated separately in Section 6 because it produces models that operate in joint image-text embedding space. CLIP (Radford et al., ICML 2021), ALIGN, OpenCLIP (LAION), SigLIP, and CoCa (Yu et al., 2022) all use ViT image encoders with at least 100M parameters trained on hundreds of millions to billions of image-text pairs.

The empirical hierarchy that has emerged by 2026 is roughly: for ImageNet fine-tuning accuracy at scale, MIM and supervised JFT pretraining are tied within 0.5%; for linear-probe and frozen-feature evaluation, DINOv2-style self-distillation is best; for zero-shot transfer to novel concepts, CLIP-style image-text contrast is the only competitive option. Practitioners increasingly combine these paradigms — EVA-02 stacks CLIP-distilled image-text features with MIM, and ViT-22B itself uses an internal sigmoid-loss SigLIP-style objective. Section 6 examines the multimodal branch in detail.

6. Vision-Language Foundation Models Built on ViT

Building on Section 5, this section turns to vision-language foundation models pairing ViT image encoders with text across three paradigms: contrastive dual towers, contrastive plus captioning, and frozen-LLM adapter models. The dual-tower paradigm includes CLIP (Radford et al., 2021, 400M-pair InfoNCE), ALIGN (Jia et al., 2021, 1.8B noisy contrastive), OpenCLIP (Cherti et al., 2022, LAION-2B), SigLIP (Zhai et al., 2023, sigmoid pairwise loss), Chinese CLIP (Yang et al., 2022, 200M Chinese pairs), and EVA-CLIP (Sun et al., 2023, MIM plus CLIP). The contrastive-plus-captioning paradigm is led by CoCa (Yu et al., 2022, contrast plus captioning decoder) and BLIP-2 (Li et al., 2023, Q-Former bridge to frozen LLM). The frozen-LLM adapter paradigm includes Flamingo (Alayrac et al., 2022, gated cross-attention) and PaLI-X (Chen et al., 2023, ViT-22B plus 32B mT5); LLaVA-1.5 (Liu et al., 2023, CLIP plus Vicuna-13B), MiniGPT-4 (Zhu et al., 2023, frozen ViT plus Vicuna), and Qwen-VL (Bai et al., 2023, multilingual VLM) extend the same template. Open-source frontiers include InternVL (2024, InternViT-6B plus LLaMA-2), IDEFICS (HuggingFace, 2023, open Flamingo), and Yi-VL (01.AI, 2024, Chinese-English VLM). Open-vocabulary recognition includes OWL-ViT (Minderer et al., 2022, open-vocab detection), Grounding DINO (Liu et al., 2023, language-grounded detection), GLIP (Li et al., 2022, grounded image-language pretraining), and SAM (Kirillov et al., 2023, promptable segmentation), while Frozen CLIP Video Learner (Lin et al., 2022) attaches a temporal head to a frozen CLIP encoder.

The most important downstream consequence of the Vision Transformer revolution has arguably been the rise of large vision-language foundation models. Because both image and language representations can be expressed as token sequences and processed by a Transformer, ViTs slot naturally into multimodal models. CLIP, ALIGN, CoCa, BLIP, BLIP-2, Flamingo, PaLI-X, LLaVA, and SAM all use ViT-family encoders to produce dense visual features that interact with language through cross-attention or contrastive alignment. This section reviews the three principal multimodal pretraining paradigms — contrastive dual-tower, contrastive-plus-captioning, and frozen-LLM adapters — and the open-vocabulary recognition benchmarks that have emerged.

6.1. CLIP, ALIGN, SigLIP and contrastive dual towers

CLIP (Contrastive Language-Image Pretraining; Radford et al., ICML 2021) consists of a ViT image encoder and a Transformer text encoder; the two encoders are trained jointly to maximize the cosine similarity of paired (image, caption) embeddings while minimizing similarity of unpaired examples. Training data is WebImageText, an internal 400M image-text pair corpus scraped from the public web. The contrastive loss is symmetric InfoNCE with temperature 0.07 (initially) annealed downward. CLIP ViT-L/14 reaches 75.5% zero-shot top-1 on ImageNet — competitive with a fully supervised ResNet-50 — and transfers strongly to 30+ downstream classification benchmarks without any fine-tuning. CLIP ViT-B/32 is the smaller and most-deployed version and has approximately 87M parameters in the image encoder alone.

Method	Objective	Pretrain data	Backbone	IN-1k FT/lin
Supervised IN-21k	Cross-entropy	14M	ViT-L/16	85.3% FT
MAE	Pixel reconstruction (75% mask)	IN-1k	ViT-H/14	87.8% FT
SimMIM	Pixel reconstruction (60% mask)	IN-1k	Swin-B	83.8% FT
BEiT	Discrete dVAE token prediction	IN-1k	ViT-L/16	86.3% FT
MaskFeat	HOG reconstruction	IN-1k	ViT-L/16	86.4% FT
DINO	Self-distillation [CLS]	IN-1k	ViT-B/16	78.2% lin
iBOT	DINO + masked tokens	IN-22k	ViT-L/16	84.0% lin
DINOv2	iBOT + curation	LVD-142M	ViT-g/14	86.5% lin
CLIP (ViT)	Image-text contrast	LAION-400M	ViT-L/14	75.5% zs
EVA-CLIP	CLIP + MIM	LAION-2B	ViT-g/14	79.0% zs

ALIGN (Jia et al., ICML 2021) followed the same architecture with a noisier 1.8B-image-text-pair corpus and demonstrated that scale of data dominates curation quality. OpenCLIP (Cherti et al., 2022) reproduced and exceeded CLIP using LAION-400M and LAION-2B (Schuhmann et al., 2022); LAION-5B contains 5.85B publicly available image-text pairs spanning 100+ languages. EVA-CLIP combined CLIP-style training with MIM and reached 79.0% zero-shot ImageNet with ViT-g/14. SigLIP (Sigmoid Loss for Language Image Pre-Training; Zhai et al., ICCV 2023) replaced the per-batch softmax of InfoNCE with a per-pair sigmoid loss, achieving better small-batch scalability and reaching 79.0% zero-shot ImageNet with ViT-L/14 trained on a 4B-pair WebLI subset. Chinese CLIP (Yang et al., 2022) extended CLIP to Chinese with a 200M-pair corpus.

Recent robustness studies (Tu et al., 2024) showed that CLIP models are remarkably robust under distribution shift: CLIP ViT-L/14 outperforms supervised ImageNet ViT-L by 25+ percentage points on ImageNet-A and ImageNet-R. However, CLIP fine-grained recognition remains imperfect — CLIP misclassifies fine-grained bird species at twice the rate of humans, and Lavoie et al. (2024) found that CLIP captures only the dominant caption interpretation of an image, missing alternative valid descriptions.

6.2. CoCa, BLIP-2, Flamingo, PaLI-X

A second multimodal paradigm augments contrastive learning with generative captioning. CoCa (Contrastive Captioner; Yu et al., 2022) decouples the image encoder into two stages: a unimodal encoder of bottom layers that produces dense features, and a multimodal decoder of top layers with cross-attention to text. The model is trained jointly with a contrastive loss between [CLS]-pooled image features and text features, and a captioning cross-entropy loss between the multimodal decoder and the caption tokens. CoCa-2B reaches 91.0% top-1 on ImageNet (with frozen-encoder MAP probe), 88.6% with fine-tuning, and 143.6 CIDEr on MS-COCO captioning, simultaneously achieving SOTA on classification, retrieval, and captioning.

BLIP-2 (Li et al., 2023) introduced the Q-Former, a small Transformer that connects a frozen ViT image encoder to a frozen large language model (LLM) like OPT or FlanT5; only the Q-Former (~108M params) is trained. This enables vision-language pretraining at much lower cost than full end-to-end training. BLIP-2 with ViT-g (1.1B) and FlanT5-XXL (11B) reaches 65.0 CIDEr on COCO captioning zero-shot.

Flamingo (Alayrac et al., NeurIPS 2022) used cross-attention adapters between a frozen NFNet-F6 image encoder (later versions used ViT) and a frozen Chinchilla LLM, enabling few-shot in-context visual question answering. PaLI-X (Chen et al., 2023) scaled this paradigm to 55B parameters with a ViT-22B image encoder and a 32B mT5 language model, training on a multilingual mixture; PaLI-X 55B is the first model to exceed human accuracy on the VTAB benchmark and scores 88.5% on VQAv2.

LLaVA-1.5 and LLaVA-1.6 (Liu et al., 2023, 2024) connected a CLIP ViT-L/14 image encoder to a Vicuna-13B language model through a two-layer MLP projector, demonstrated that a small instruction-tuning dataset (~150K conversations from GPT-4) was sufficient to produce a capable multimodal chatbot, and made open-source MLLM development practical. A growing taxonomy of open-source vision-language models — MiniGPT-4, Qwen-VL, InternVL, IDEFICS, Yi-VL — all follow the same template: ViT image encoder + projection + LLM decoder.

6.3. Open-vocabulary recognition and zero-shot transfer

Vision-language ViTs excel at open-vocabulary recognition — recognizing objects from arbitrary text de-

scriptions at inference time, without any fine-tuning. CLIP's most striking demonstration was zero-shot ImageNet at 75.5% top-1 with no ImageNet labels used during training. Subsequent work has extended this capability to detection, segmentation, and even point clouds. CRIS (Wang et al., CVPR 2022) is a CLIP-driven referring image segmentation framework. Open-vocabulary detection methods like OWL-ViT, GLIP, and Grounding DINO use CLIP-style image-text alignment to detect novel categories. CRIS, MaskCLIP, OpenVocabCT (Li et al., 2025), and Self-Calibrated CLIP (Bai et al., 2025) extend CLIP to dense pixel labeling.

Frozen CLIP models are also strong video learners. Frozen CLIP Models are Efficient Video Learners (Lin et al., ECCV 2022) demonstrated that a frozen CLIP ViT-L/14 plus a small temporal Transformer head reaches state-of-the-art on Kinetics-400 with a fraction of the training cost. Wu et al. (AAAI 2023) revisited the CLIP-as-video-classifier idea and reached 85.4% on K-400.

CLIP and its descendants have also been adapted to specialized domains. RadCLIP (Lu et al., 2025) and BiomedCLIP target radiology; AquaticCLIP (Alawode et al., 2026) targets underwater scenes; DOFA-CLIP (Xiong et al., 2025) targets Earth observation; PestCLIP targets agriculture. CLIP in Medical Imaging: A Survey (Zhao et al., 2025, Medical Image Analysis) inventories 30+ CLIP-derived medical models.

By 2026, the line between "vision Transformer" and "vision-language foundation model" has blurred. Almost all serious ViT pretraining now incorporates some form of language supervision, either contrastively (CLIP, SigLIP) or generatively (CoCa, BLIP-2). The downstream consequence is that classification, detection, and segmentation are increasingly handled by a single multimodal model exposed via prompts, rather than by task-specific fine-tuning. The CLIP-as-feature-extractor paradigm has effectively reached parity with DINOv2 for many transfer tasks. Section 7 next describes how dense-prediction transformers — including the rapidly evolving DETR family — have been refined alongside this multimodal trajectory.

7. Dense Prediction with Transformers: Detection, Segmentation, and Tracking

Whereas Section 6 covered vision-language foundations, this section turns to dense prediction with Transformer decoders across three families: the DETR detection family, segmentation Transformers, and task-specialized backbones. The DETR family includes DETR (Carion et al., 2020, set prediction with Hungarian matching), Deformable DETR (Zhu et al., 2021, k=4 reference-point sampling), Conditional DETR (Meng et al., 2021, decoupled content-position queries), DAB-DETR (Liu et al., 2022, 4D anchor-box queries), DN-DETR (Li et al., 2022, denoising training), DINO-DETR (Zhang et al., 2023, contrastive denoising plus mixed selection), Co-DETR (Zong et al., 2024, one-to-many auxiliary matching, 66.0 AP), and RT-DETR (Zhao et al., 2024, real-time hybrid encoder, 53.0 AP at 114 FPS), and extends to MOTR (Zeng et al., 2022, DETR plus track queries) and OW-DETR (Gupta et al., 2022, open-world detection). Segmentation Transformers include SETR (Zheng et al., 2021, plain ViT segmentation), SegFormer (Xie et al., 2021, hierarchical MiT plus all-MLP decoder), Segmenter (Strudel et al., 2021, class-mask tokens), MaskFormer (Cheng et al., 2021, mask classification), Mask2Former (Cheng et al., 2022, masked attention plus multi-scale, 57.7 mIoU), OneFormer (Jain et al., 2023, task-token unified segmenter), and K-Net (Zhang et al., 2021, dynamic kernels). Task-specialized backbones include ViT-Det (Li et al., 2022, plain-ViT single-scale FPN, 61.3 AP), ViT-Adapter (Chen et al., 2023, lightweight spatial adapter, 60.5 mIoU), BEVFormer (Li et al., 2022, deformable BEV cross-attention, 51.7 NDS), and PETR (Liu et al., 2022, 3D-position-encoded queries).

Dense prediction tasks include object detection, instance and semantic segmentation, panoptic segmentation, depth estimation, optical flow, and tracking. They were historically dominated by convolutional architectures with feature pyramids (FPN), region-of-interest pooling (RoI), and non-maximum suppression (NMS). Vision Transformers have now largely displaced CNNs in this regime. Two factors drove the shift. First, hierarchical ViTs (Swin, MViT, PVT) provide better backbones than ResNets at matched FLOP budgets. Second, Transformer-based decoders (DETR, MaskFormer, Mask2Former) replace many hand-designed components with set prediction. This section reviews the three currently dominant families.

7.1. DETR family: DETR, Deformable DETR, DINO-DETR, RT-DETR

DETR (Carion et al., ECCV 2020) recasts object detection as a direct set-prediction problem. A CNN or ViT backbone extracts features, a Transformer encoder refines them, and a Transformer decoder takes N learnable object queries (typically $N = 100$) plus encoder memory and outputs N box-class predictions. Bipartite matching via the Hungarian algorithm assigns predictions to ground-truth boxes, after which a

Model	Image encoder	Text/LM	Pretraining data	Headline metric
CLIP (Radford 2021)	ViT-L/14 (304M)	TextEncoder-12L	WIT-400M	75.5% IN-1k zero-shot
ALIGN (Jia 2021)	EfficientNet-L2 + later ViT	BERT	1.8B noisy	76.4% IN-1k zs
OpenCLIP (Cherti 2022)	ViT-H/14	TextEncoder	LAION-2B	78.0% IN-1k zs
SigLIP (Zhai 2023)	ViT-L/14	TextEncoder	WebLI 4B	79.0% IN-1k zs
CoCa (Yu 2022)	ViT-G	Decoder + contrast	ALIGN+JFT	91.0% IN-1k MAP
BLIP-2 (Li 2023)	ViT-g + Q-Former	FlanT5 / OPT	129M	65 CIDEr COCO zs
Flamingo (Alayrac 2022)	NFNet-F6 / ViT	Chinchilla 80B	M3W + ALIGN	56.0% VQAv2 4-shot
PaLI-X (Chen 2023)	ViT-22B	mT5-XXL	WebLI multilingual	88.5% VQAv2
LLaVA-1.5 (Liu 2023)	CLIP ViT-L/14	Vicuna-13B	LLaVA-Instruct-150K	80.0% ScienceQA
EVA-CLIP (Sun 2023)	ViT-g/14	TextEncoder	LAION-2B + MIM	79.0% IN-1k zs
InternVL (2024)	InternViT-6B	LLaMA-2	LAION + COYO	80.7% MMBench
SAM (Kirillov 2023)	ViT-H	Prompt encoder	SA-1B (1.1B masks)	Zero-shot segmentation

combined loss of focal classification, ℓ_1 box, and GIoU box is back-propagated. DETR with a ResNet-50 backbone reaches 42.0 box AP on COCO val2017, and DETR-DC5-R50 (with dilated convolutions) reaches 43.3 AP. The original DETR converges slowly — 500 epochs on COCO with a peak LR of 1e-4 — because the cross-attention from queries to encoded features must learn from scratch where to look.

Deformable DETR (Zhu et al., ICLR 2021) replaced full cross-attention by a deformable variant that samples $k = 4$ reference points per query head. This dropped epochs from 500 to 50 and raised AP to 49.5 with ResNet-50 and 52.3 with Swin-T. Conditional DETR added conditional spatial queries that decouple content and position. DAB-DETR introduced 4D anchor boxes as queries. DN-DETR (denoising DETR) adds a denoising group of noised ground-truth queries during training to stabilize bipartite matching. DINO-DETR (Zhang et al., 2023) combined contrastive denoising, mixed query selection, and look-forward-twice and reached 63.3 box AP on COCO with Swin-L, the SOTA for 2023.

Co-DETR (Zong et al., 2024) further pushes accuracy by combining one-to-many matching during training with one-to-one matching during inference, reaching 66.0 AP. RT-DETR (Zhao et al., CVPR 2024) is the first real-time detector that beats YOLO at all common operating points: RT-DETR-L (32M params) reaches 53.0 AP at 114 FPS on T4, and RT-DETR-X (67M) reaches 54.8 AP at 74 FPS. The core insight of RT-DETR is hybrid encoders that decouple intra-scale interaction from cross-scale fusion, plus IoU-aware query selection to seed the decoder. A recent comprehensive survey by Yu et al. (2025) catalogs more than 30 DETR variants.

DETR ideas have been extended beyond detection. MOTR (Zeng et al., ECCV 2022) adds track queries to DETR for multi-object tracking, achieving 67.4 IDF1 on MOT17. OW-DETR (Gupta et al., CVPR 2022) extends DETR to open-world detection. TransVOD (Zhou et al., TPAMI 2023) processes video as a temporal sequence of object queries.

7.2. Segmentation transformers: SETR, SegFormer, MaskFormer, Mask2Former

SETR (Zheng et al., CVPR 2021) was the first pure-Transformer semantic segmentation network. A ViT-Large/16 encoder produces a fixed-resolution feature map; a simple progressive upsampling decoder restores pixel resolution. SETR-PUP with ViT-L reaches 50.3 mIoU on ADE20K, demonstrating that even without any inductive locality bias, a ViT can match the contemporaneous DeepLabV3+ baseline.

SegFormer (Xie et al., NeurIPS 2021) is the most influential segmentation Transformer to date. It introduces a hierarchical Mix Transformer (MiT) en-

coder with overlapping patch merging and efficient self-attention (similar to PVT's spatial-reduction attention) and an extremely lightweight all-MLP decoder. SegFormer scales from MiT-B0 (3.8M params, 8.4 GFLOPs) reaching 37.4 mIoU on ADE20K, to MiT-B5 (84M params, 183 GFLOPs) reaching 51.0 mIoU on ADE20K and 84.0% mIoU on Cityscapes — at the time, SOTA at any FLOP budget. SegFormer's small models are still the dominant choice for real-time semantic segmentation in 2026.

MaskFormer (Cheng et al., NeurIPS 2021) reformulated semantic segmentation as a mask-classification problem rather than per-pixel classification. A Transformer decoder produces $N = 100$ binary masks plus class predictions; semantic segmentation becomes a special case in which class prediction collapses to one prediction per category. Mask2Former (Cheng et al., CVPR 2022) added masked attention — confining cross-attention to currently predicted foreground regions — and multi-scale features, becoming a unified architecture for semantic, instance, and panoptic segmentation. Mask2Former with Swin-L backbone reaches 57.7 mIoU on ADE20K, 50.5 PQ on COCO panoptic, and 50.1 AP on COCO instance, simultaneously delivering SOTA on all three tasks. It remains the strongest segmentation architecture in 2024 outside of Segment Anything-style foundation models.

Other notable segmentation transformers include Segmenter (Strudel et al., ICCV 2021) which uses class-mask tokens; K-Net which unifies instance/semantic/panoptic via dynamic convolutional kernels; OneFormer which conditions on a task token to handle three segmentation tasks with one model; Point Transformer V3 Extreme which won the 2024 Waymo semantic segmentation challenge; and Swin-Unet (Cao et al., 2023), Swin UNETR (Hatamizadeh et al., 2022), TransDeepLab, TransUNet, and UCTransNet — all medical adaptations.

7.3. Specialized backbones: ViT-Det, ViT-Adapter, BEVFormer

A late development in dense prediction has been the rehabilitation of the plain ViT backbone for tasks that traditionally required hierarchical features. ViTDet (Li et al., ECCV 2022) showed that with a single high-resolution feature map produced by a plain ViT encoder, plus a simple feature pyramid built by striding/transposing/dilating that single map, plain ViTs match or exceed Swin-based detectors. ViT-Det with MAE-pretrained ViT-H reaches 61.3 box AP on COCO. ViT-Adapter (Chen et al., ICLR 2023) injects spatial priors into a frozen ViT through a lightweight adapter network that shares features with the ViT at multiple scales, achieving 60.5 mIoU on ADE20K with ViT-Adapter-L and BEiT pretraining, matching Swin V2-G with one-fifth the parameters.

For autonomous driving, BEVFormer (Li et al., ECCV 2022) constructs a bird's-eye-view (BEV) feature representation by querying multi-camera image features through deformable cross-attention. BEVFormer reaches 51.7% NDS on nuScenes, beating all previous LiDAR-free methods. PETR (Liu et al., ECCV 2022) and PETR v2 use a similar query-based scheme but with 3D positional encoding embedded in the camera features. DETR3D extends DETR object queries to 3D space. V2X-ViT (Xu et al., ECCV 2022) extends BEV processing to vehicle-to-everything cooperative perception, sharing features between an ego vehicle and roadside units.

The cumulative effect of these advances is that by 2026, almost every leaderboard for dense prediction is led by a Transformer-based model, often combining a ViT backbone, a DETR-style decoder, and pretraining on a foundation-scale corpus. The remaining battlegrounds are real-time inference (where RT-DETR and YOLO-derived hybrids compete), 3D scene understanding (where BEV-based ViTs co-exist with LiDAR-specialized point Transformers), and small-object detection (where Deformable attention's reference-point sampling still struggles below 32×32 pixel targets). Section 8 turns to spatio-temporal and 3D extensions.

8. Spatio-Temporal and 3D Transformers

Building on Section 7, this section turns to ViTs for video and 3D data across three families: video classification, promptable video segmentation, and point-cloud Transformers. Video classification includes TimeSformer (Bertasius et al., 2021, divided space-time, 82.1% K-400), ViViT (Arnab et al., 2021, factorized encoder Model 2, 84.9% K-400), MViT v1/v2 (Fan et al., 2021/2022, hierarchical pooling, 88.8% K-400), Video Swin (Liu et al., 2022, 3D shifted windows, 84.9% K-400), VideoMAE (Tong et al., 2022, 90% tube masking, 87.4% K-400), VideoMAE V2 (Wang et al., 2023, ViT-g, 90.0% K-400), MAE-ST (He et al., 2022, spatio-temporal MAE), Hiera (Ryali et al., 2023, MAE-distilled, 86.8% K-400), MaskFeat (Wei et al., 2022, video HOG), and Action Transformer (Mazzia et al., 2021, pose-based action), with EgoVLP (Lin et al., 2022, egocentric VLP) and InternVideo (Wang et al., 2023, 1.4B-parameter video foundation). Promptable video segmentation is led by SAM 2 (Ravi et al., 2024, Hiera plus memory bank, 79.5 J&F). Track-

Task	Best transformer model	Backbone	Benchmark	Headline metric
COCO detection	Co-DETR	Swin-L	COCO val2017	66.0 box AP
COCO detection	DINO-DETR	Swin-L	COCO val2017	63.3 box AP
COCO detection (real-time)	RT-DETR-X	HGNet-X	COCO @114 FPS	54.8 box AP
ADE20K semantic	Mask2Former	Swin-L	ADE20K	57.7 mIoU
ADE20K semantic	SegFormer-B5	MiT-B5	ADE20K	51.0 mIoU
ADE20K semantic	ViT-Adapter-L	BEiT-L	ADE20K	60.5 mIoU
Cityscapes	SegFormer-B5	MiT-B5	Cityscapes	84.0 mIoU
COCO panoptic	Mask2Former	Swin-L	COCO panoptic	50.5 PQ
MOT17 tracking	MOTR	ResNet-50	MOT17 test	67.4 IDF1
nuScenes 3D detection	BEVFormer	ResNet-101	nuScenes test	51.7 NDS
Open-vocab detection	OWL-ViT	ViT-L/14	LVIS rare	31.2 AP
Open-vocab segmentation	OpenSeeD	Swin-L	ADE20K (zero-shot)	23.4 mIoU

ing includes MOTRv2 (Zhang et al., 2023, anchor-conditioned), TransTrack (Sun et al., 2021, query-based), and TrackFormer (Meinhardt et al., 2022, joint detection-tracking queries), and Mask2Former-VIS (Cheng et al., 2021) extends masked attention to video instance segmentation. Point-cloud Transformers include PCT (Guo et al., 2021, offset attention, 93.2% ModelNet40), Point Transformer V1/V2 (Zhao et al., 2021/2022, vector self-attention, 93.7%), Point-MAE (Pang et al., 2022, point-cloud MIM, 90.0% ScanObjectNN), Inter-Modal MAE (Liu et al., 2023, cross-modal point SSL), and Point Transformer V3 (Wu et al., 2024, serialized 1D Transformer, 73.5 ScanNet mIoU), with V3 Extreme winning the 2024 Waymo challenge. Vim4Path (Nasiri-Sarvi et al., 2024) applies Vision Mamba to pathology and SRT (Sajjadi et al., 2022) is a scene representation Transformer for novel-view synthesis.

Vision Transformers extend naturally to data with additional temporal or geometric structure. Video, point clouds, multi-view images, and event-camera streams can all be tokenized into sequences and processed with the same Transformer skeleton. The literature has produced a rich set of design choices — joint vs. factorized space-time attention, asymmetric masking ratios, attention with positional encodings on irregular point grids — that we review here.

8.1. Video classification: TimeSformer, ViViT, MViT v2, VideoMAE, Hiera

The simplest video Transformer treats each video clip as a sequence of $T \times H/P \times W/P$ tokens, where T is the number of frames. With $T = 8$ and 224^2 resolution at patch size 16, this gives $N = 8 \times 14 \times 14 = 1568$ tokens — eight times more than a single image — which makes joint space-time attention prohibitively expensive at $N^2 = 2.46 \times 10^6$ pairs per layer.

ViViT (A Video Vision Transformer; Arnab et al., ICCV 2021) systematically explored four factorization schemes: (1) spatio-temporal joint attention; (2) factorized encoder (spatial then temporal); (3) factorized self-attention (alternating spatial and temporal); (4) factorized dot product (separate spatial and temporal heads). The factorized encoder, called Model 2 in the paper, is the strongest and reaches 84.9% top-1 on Kinetics-400 with ViViT-L. TimeSformer (Bertasius et al., ICML 2021) introduced "divided space-time attention" — in the same architecture, alternating layers attend over space and time — and reached 82.1% on K-400. Multiscale Vision Transformers v1/v2 (Fan et al., ICCV 2021/CVPR 2022) combined hierarchical pyramidal structure with spatio-temporal pooling attention, reducing the number of tokens at deeper stages by adaptive 3D pooling. MViTv2-L reaches 88.8% on K-400 — SOTA at the time. Video Swin Transformer (Liu et al., 2022) extended Swin's shifted windows to 3D, reaching 84.9% on K-400 and 86.8% on K-600.

VideoMAE (Tong et al., NeurIPS 2022) extended MAE to video. The asymmetric encoder-decoder design uses a tube masking strategy at mask ratio 90% (much higher than image MAE's 75%) to exploit the high redundancy of consecutive frames. VideoMAE pretrained on Kinetics-400 reaches 87.4% top-1 fine-tuning with ViT-L; MAE-ST (He et al., 2022) achieves similar results on action recognition benchmarks. VideoMAE V2 scales the model to ViT-g and reaches 90.0% on K-400 and 71.6% on Something-Something V2. Hiera (Ryali et al., ICML 2023) is a hierarchical ViT distilled from MAE pretraining on video that simplifies the MViT architecture, remov-

ing all the conv-based pooling, and reaches 86.8% on K-400 with fewer parameters.

For action detection, MViT and TimeSformer feature into AVA v2.2 detectors. EgoVLP (Egocentric Video-Language Pretraining) and InternVideo (1.4B params) push the multimodal frontier on Ego4D, EPIC-Kitchens-100 (4M action segments), and Something-Something V2 (220K clips). Action Transformer (Mazzia et al., 2021) handles short-time pose-based action recognition.

8.2. Promptable video segmentation: SAM 2 and Hiera variants

In August 2024, Meta released SAM 2 (Ravi et al.), the segment-anything model extended to video. SAM 2 uses a Hiera-S/B+/L image encoder, a memory bank that stores per-frame features and per-frame mask tokens, a memory attention module that conditions current-frame predictions on past memory, and a mask decoder identical to SAM. The training data is SA-V, a 50.9-K-video, 35.5-million-mask dataset. SAM 2-L reaches 79.5 J&F on YouTube-VOS 2018 and 78.1 on DAVIS 2017 in zero-shot mode, surpassing previous task-specific models. Recent fine-tuning work like SurgiSAM2 and SAM-I2V++ adapts SAM 2 to specialized domains such as surgical video and biomedical microscopy.

Other video Transformers for tracking include MOTR (DETR-based MOT), MOTRv2, TransTrack, TrackFormer, and OmniMOTION; for video object segmentation, MaskFreeVIS, Mask2Former for VIS (Cheng et al., 2021), and the Generic Video Transformer family. For video instance segmentation on YouTube-VIS 2019, Mask2Former for VIS reaches 55.1 AP with Swin-L.

8.3. Point clouds and 3D scenes: PCT, Point-MAE, Point Transformer V3

3D point clouds present a different tokenization challenge: points are unordered and irregularly sampled, so there is no natural patch grid. Point Cloud Transformer (PCT; Guo et al., CVMJ 2021) used input embeddings from k-NN neighborhoods plus offset attention (computing attention from differences rather than raw values), achieving 93.2% accuracy on ModelNet40. Point Transformer V1 (Zhao et al., ICCV 2021) introduced vector self-attention, which uses subtraction-then-MLP rather than dot-product, and reached 93.7% on ModelNet40 and 70.4 mIoU on S3DIS Area 5.

Point-MAE (Pang et al., ECCV 2022) ports MAE to point clouds by sampling a set of point patches via FPS+kNN, masking 60% of patches, and reconstructing them with a Chamfer distance loss. Point-MAE pretrained on ShapeNet (51,300 shapes) and fine-tuned on ScanObjectNN reaches 90.0% (PB_T50_RS hardest variant), a substantial gain over training from scratch. Inter-Modal MAE (Liu et al., 2023) and the contrastive variant in Bringing MAE Explicit Contrastive Properties (Ren et al., 2024) further explore point-cloud SSL.

Point Transformer V3 (Wu et al., CVPR 2024) and its V3 Extreme version (Wu et al., 2024) won the 2024 Waymo Open Dataset Challenge in semantic segmentation. PTv3 simplifies the architecture by removing complex local feature aggregation, instead relying on serialized point sequences — Hilbert-curve and Z-order space-filling curves convert 3D coordinates to 1D tokens that are processed by efficient 1D Transformers. PTv3 reaches 73.5 mIoU on ScanNet v2 and 89.1% on ScanObjectNN, setting state-of-the-art at substantially lower compute than V2.

For multi-view 3D understanding, NeRF-related Transformers like SRT (Scene Representation Transformer) and IBRNet condition novel-view synthesis on a few input images. For surgical video and medical 3D imaging, Vim4Path (Nasiri-Sarvi et al., 2024) applies Vision Mamba to gigapixel whole-slide histopathology, demonstrating that Mamba-based 3D models can outperform ViT-based foundation models like UNI on downstream slide-level prediction.

The two systemic lessons from spatio-temporal and 3D ViTs are: (i) factorization or selection (TimeSformer's divided attention, MViT's pooling, Mask2Former's masked attention, SAM 2's memory bank) is essential for scaling beyond single-image tokens; (ii) self-supervised reconstruction with high mask ratios (75% for image MAE, 90% for VideoMAE, 60% for Point-MAE) is strikingly effective across modalities, suggesting that the asymmetric encoder-decoder template generalizes well beyond pixels. Section 9 turns to the datasets and benchmarks that anchor these evaluations.

9. Datasets, Benchmarks, and Evaluation Protocols

Whereas Section 8 reviewed video and 3D models, this section turns to the datasets and benchmarks against which ViTs are measured across three groupings: pretraining corpora, downstream task benchmarks, and robustness suites. Pretraining corpora include ImageNet-1k (Deng et al., 2009, 1.28M, 1000

Domain	Method	Year	Architecture	Pretraining	Headline metric
Video classification	TimeSformer	2021	Divided space-time attn	IN-21k	82.1% K-400
Video classification	ViViT (Model 2)	2021	Factorized encoder	IN-21k	84.9% K-400
Video classification	MViT v2-L	2022	Hierarchical 3D	K-400	88.8% K-400
Video classification	Video Swin-L	2022	3D Swin	IN-21k	84.9% K-400
Video classification	VideoMAE-L	2022	MAE on tubes	K-400 self-sup	87.4% K-400
Video classification	VideoMAE V2-g	2023	MAE	K-700 self-sup	90.0% K-400
Video classification	Hiera-H	2023	Hierarchical, MAE-distill	K-400	86.8% K-400
Video segmentation	SAM 2-L	2024	Hiera + memory	SA-V (50.9K vids)	79.5 J&F YT-VOS
Video instance seg	Mask2Former-VIS	2021	Mask2Former	IN-1k	55.1 AP YT-VIS19
Multi-object tracking	MOTR	2022	DETR + track queries	COCO	67.4 IDF1 MOT17
Point cloud cls	PCT	2021	Offset attention	n/a	93.2% ModelNet40
Point cloud cls	Point Transformer V1	2021	Vector self-attn	n/a	93.7% ModelNet40
Point cloud SSL	Point-MAE	2022	MAE for points	ShapeNet	90.0% ScanObjectNN
Point cloud seg	Point Transformer V3	2024	Serialized 1D Transformer	n/a	73.5 ScanNet mIoU
Pathology WSI	Vim4Path	2024	Vision Mamba	TCGA	beats UNI on slide pred

classes), ImageNet-21k (Deng et al., 2009, 14M, 21,841 classes), JFT-300M (Sun et al., 2017, 300M weakly labeled), JFT-3B (Zhai et al., 2022, 3B), LAION-400M / LAION-2B / LAION-5B (Schuhmann et al., 2022, 5.85B image-text pairs), WebLI (Chen et al., 2023, 12B multilingual), DataComp-12M (Gadre et al., 2023, curated CLIP benchmark), CC12M (Changpinyo et al., 2021, 12M conceptual captions), LVD-142M (Oquab et al., 2023, DINOv2-curated), SA-1B (Kirillov et al., 2023, 11M images, 1.1B masks), and SA-V (Ravi et al., 2024, 50.9K videos, 35.5M masks). Downstream benchmarks include COCO 2017 (Lin et al., 2014, 118K det/seg), LVIS v1 (Gupta et al., 2019, long-tail detection), ADE20K (Zhou et al., 2017, 25K scene parsing), Cityscapes (Cordts et al., 2016, 5K urban), Kinetics-400/600/700 (Kay et al., 2017), Something-Something V2 (Goyal et al., 2017, 220K motion clips), AVA v2.2 (Gu et al., 2018), EPIC-Kitchens-100 (Damen et al., 2022), nuScenes (Caesar et al., 2020), Waymo Open (Sun et al., 2020), ModelNet40 (Wu et al., 2015), ScanNet v2 (Dai et al., 2017), VQAv2 (Goyal et al., 2017), MS-COCO Captions (Chen et al., 2015), RefCOCO (Yu et al., 2016), YouTube-VOS 2018 (Xu et al., 2018), and DAVIS 2017 (Pont-Tuset et al., 2017). Robustness suites include ImageNet-V2 (Recht et al., 2019, replication shift), ImageNet-A (Hendrycks et al., 2021, natural adversarial), ImageNet-R (Hendrycks et al., 2021, renditions), ImageNet-Sketch (Wang et al., 2019, sketches), ImageNet-C (Hendrycks and Dietterich, 2019, 19 corruptions), ObjectNet (Barbu et al., 2019, viewpoint shift), DomainNet (Peng et al., 2019, six-domain shift), and WILDS (Koh et al., 2021, real-world shift).

A faithful evaluation of Vision Transformers requires consistent datasets, well-defined metrics, and explicit reporting of pretraining corpora. This section catalogs the most important datasets and benchmarks used by the ViT literature, organized by task family, and discusses the evaluation protocols that underpin the numbers cited throughout this survey.

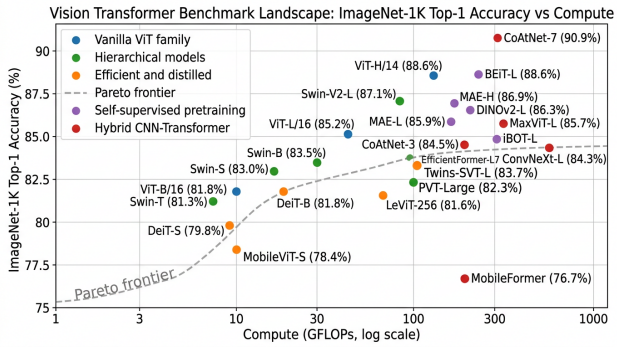

Figure 5. Benchmark landscape

9.1. Pretraining corpora: ImageNet-21k, JFT-3B, LAION-5B, SA-1B

The single largest determinant of a ViT's downstream performance is the pretraining corpus. Five corpora dominate.

ImageNet-1k (Deng et al., CVPR 2009) contains 1.28 million training images, 50,000 validation images, and 100,000 test images organized into 1000 mutually exclusive object categories. It is the standard fine-tuning and evaluation benchmark, and after the introduction of MAE in 2022 it also became a viable pretraining corpus on its own.

ImageNet-21k is the full version of ImageNet, containing 14,197,122 images across 21,841 WordNet synsets. It is the largest publicly available image classification corpus and is used to pretrain ViT-B, ViT-L, ViT-H, Swin-L, and ConvNeXt-XL backbones. Steiner et al. (2022) showed that with the right augmentation recipe IN-21k pretraining matches JFT-300M for ViT-B and ViT-L.

JFT-300M and JFT-3B are Google-internal corpora of 300 million and 3 billion weakly labeled images respectively. ViT, ViT-G, ViT-22B, CoAtNet-7, and PaLI-X all use JFT for pretraining. JFT is not publicly available, which has motivated the open-source LAION effort.

LAION-400M, LAION-2B, and LAION-5B (Schuhmann et al., 2022) are open-source web-scraped image-text pair datasets. LAION-5B contains 5.85 billion pairs in 100+ languages, of which 2.32B are English. OpenCLIP, EVA-CLIP, and Stable Diffusion all train on LAION subsets. The corpus has known issues — dataset cleanup removed some content in 2024 — but remains the largest open multimodal corpus.

SA-1B (Kirillov et al., 2023) is a segmentation dataset created during SAM training: 11.1 million images annotated with 1.1 billion automatically generated, then human-quality-controlled, segmentation masks. SA-1B is two orders of magnitude larger than COCO segmentation and enables training of segmentation foundation models that generalize zero-shot.

Other notable pretraining corpora include WebLI (used by PaLI and SigLIP, 12B image-text pairs in 109 languages), DataComp-12M (Gadre et al., 2023, a curated CLIP benchmark), Conceptual Captions 12M (CC12M), and the LVD-142M corpus curated by DINOv2 from a 1.2B-image deduplicated pool.

9.2. Downstream benchmarks for classification, detection, segmentation, video

The metrics warrant brief explanation. Top-1 accuracy on classification reports the fraction of images whose top predicted class matches the ground-truth label. Box AP (used for COCO detection) is the mean average precision averaged over IoU thresholds 0.50 to 0.95 in steps of 0.05; AP50 alone is a more permissive metric. mIoU is the mean of class-wise intersection-over-union over all categories. PQ (panoptic quality) decomposes into segmentation quality (SQ) and recognition quality (RQ): $\mathrm{PQ} = \mathrm{SQ} \cdot \mathrm{RQ}$. NDS (nuScenes detection score) is a weighted average of mAP and several true-positive errors. CIDEr and BLEU are caption-text similarity metrics. J&F in video segmentation is the average of region-similarity Jaccard $\mathcal{J}$ and contour-similarity $\mathcal{F}$.

9.3. Robustness and distribution-shift suites

Vision Transformers' robustness has become an evaluation pillar of its own. The standard suite includes:

- ImageNet-V2 (Recht et al., 2019): a re-collected ImageNet validation set with 10,000 images across the same 1000 classes, exposing benchmark overfitting. ViTs see top-1 accuracy drops of 7–12 percentage points relative to ImageNet-1k val.
- ImageNet-A (Hendrycks et al., 2021): 7,500 natural adversarial images that fool ResNet-50; the average ResNet-50 accuracy is 0%, while CLIP ViT-L/14 reaches 70.7% and ViT-22B exceeds 80%.
- ImageNet-R (Hendrycks et al., 2021): 30,000 images of 200 ImageNet classes rendered as paintings, embroidery, origami, etc. ViT-L pretrained on JFT reaches 84%, vs. ResNet-152 at 49%.
- ImageNet-Sketch (Wang et al., 2019): 50,000 black-and-white sketch images of the same 1000 ImageNet classes. CLIP ViT-L/14 zero-shot reaches 60% top-1; supervised ViT-L drops to 31%.

Benchmark	Task	Train / Val	Metric	Top-1 ViT score
ImageNet-1k	Classification	1.28M / 50k	Top-1 acc	89.5% (ViT-22B)
ImageNet-V2	Classification	— / 10k	Top-1 acc	81.6% (ViT-L)
iNaturalist 2018	Fine-grained cls	437k / 24k	Top-1 acc	87.4% (ViT-L)
Places365	Scene cls	1.8M / 36.5k	Top-1 acc	60.5% (ViT-B)
COCO 2017 (det)	Det/Inst seg	118k / 5k	box AP, mask AP	66.0 box AP (Co-DETR)
LVIS v1	Long-tail det	100k / 19.8k	AP_rare	47.3 (ViTDet-H)
ADE20K	Sem seg	25k / 2k	mIoU	60.5 (ViT-Adapter-L)
Cityscapes	Sem seg	2975 / 500	mIoU	84.0% (SegFormer-B5)
COCO panoptic	Panoptic	118k / 5k	PQ	50.5 (Mask2Former-Swin-L)
Kinetics-400	Action rec	240k / 20k	Top-1 acc	90.0% (VideoMAE V2)
Kinetics-600	Action rec	380k / 30k	Top-1 acc	88.6% (MViTv2)
Something-Something V2	Action rec	169k / 25k	Top-1 acc	75.5% (VideoMAE V2)
AVA v2.2	Action det	235k / 64k	mAP	39.1 (ACAR-Net + MViT)
EPIC-Kitchens-100	Action rec	67k / 9.7k	Verb/Noun	70.7 / 56.4 (MViTv2)
nuScenes	3D det	28k / 6k	NDS	51.7 (BEVFormer)
Waymo Open	3D seg	798 / 202	mIoU	72.5 (PTv3 Extreme)
ModelNet40	Point cls	9843 / 2468	OA	93.7% (PT V1)
ScanNet v2	3D seg	1201 / 312	mIoU	73.5 (PTv3)
VQAv2	VQA	658k / 214k	Accuracy	88.5% (PaLI-X)
MS-COCO captions	Captioning	414k / 25k	CIDEr	145+ (CoCa)
RefCOCO	Referring	142k / 50k	IoU	79.5 (Grounding DINO)
YouTube-VOS 2018	Video seg	3471 / 474	J&F	79.5 (SAM 2-L)
DAVIS 2017	Video seg	60 / 30	J&F	78.1 (SAM 2-L)

- ImageNet-C (Hendrycks and Dietterich, ICLR 2019): 19 corruption types (Gaussian noise, motion blur, fog, JPEG, etc.) at 5 severity levels. The mean corruption error (mCE) for ViT-L is around 47 vs. 76 for ResNet-50.
- ObjectNet (Barbu et al., NeurIPS 2019): 50,000 photos of objects in unusual viewpoints; designed to expose viewpoint and rotation overfitting.
- DomainNet and WILDS: domain-adaptation benchmarks across art-photo-sketch and real-world distribution shifts.

Empirically, Paul and Chen (AAAI 2022) reported that ViT-L outperforms ResNet-152 by 8–15 percentage points on ImageNet-A/R/Sketch under matched pretraining, and that the gap widens with model size. CLIP-pretrained ViTs are particularly robust on natural shifts because their text supervision is less correlated with ImageNet category boundaries. However, ViTs are not uniformly more robust: Bai et al. (2021) showed that under strong adversarial attacks (PGD-ℓ_∞ with $\varepsilon = 4/255$, 10 steps), ViT-B/16 and ResNet-50 both drop to <1% accuracy; the modest natural advantage of ViTs disappears under worst-case perturbations.

Several Transformer-specific robustness benchmarks have emerged. The Robust Vision Challenge 2022 (RVC) used ViT-Adapter-L plus Mask2Former to win the ECCV 2022 segmentation track. Large-scale Robustness Analysis of Video Action Recognition Models (Schiappa et al., 2022) systematically perturbed K-400 with 90 corruption types and observed that VideoMAE was 5–8% more robust than supervised TimeSformer. For VLM robustness, Tu et al. (2024) constructed a holistic CLIP robustness suite covering visual, textual, and modality-bridging perturbations.

The protocols of evaluation matter as much as the benchmark choice. Linear probing (LP) freezes the backbone and trains only a linear head; fine-tuning (FT) updates all weights; k-NN classification uses cosine similarity in feature space; zero-shot (ZS) evaluates without any task-specific training. DINOv2 reports all four protocols on every benchmark; CLIP reports primarily zero-shot. When comparing methods across this survey we have explicitly noted the protocol next to each number.

The cumulative consequence is that the ViT field has accumulated a richly structured benchmark stack, ranging from clean-domain accuracy (ImageNet, COCO, K-400) through distribution shift (IN-A/R/Sketch/C) to adversarial worst-case (PGD, AutoAttack). Section 10 next interprets the most important findings about how performance scales with compute and parameters.

10. Scaling Laws, Compute, and Frontier Systems

Building on Section 9, this section turns to scaling behavior and frontier systems across four parts: empirical scaling, mixture-of-experts conditional compute, training-cost budgets, and design-time lessons from scaling laws. The dense frontier includes ViT-G/14 (Zhai et al., 2022, 1.84B, 90.45% IN-1k FT), ViT-22B (Dehghani et al., 2023, 21.7B, 89.51% lin), EVA-02 (Sun et al., 2023, 304M with CLIP distillation, 90.0% IN-1k), EVA-CLIP (Sun et al., 2023, 5B contrastive, 79.0% zero-shot), DINOv2 ViT-g/14 (Oquab et al., 2023, 1.1B self-distillation, 86.5% lin), CoCa-2B (Yu et al., 2022, contrast plus captioning, 91.0% MAP), and PaLI-X (Chen et al., 2023, 55B multimodal, 88.5% VQAv2). Mixture-of-experts work includes V-MoE-15B (Riquelme et al., 2021, vision MoE, 90.35% transfer), LIMoE (Mustafa et al., 2022, joint vision-language MoE), Soft MoE (Puigcerver et al., 2024, differentiable soft routing), and Switch ViT (Fedus et al., 2022, sparse-routed Transformer). Conditional-compute approaches include DynamicViT (Rao et al., 2021, learned token dropping), Evo-ViT (Xu et al., 2022, slow-fast token evolution), and AdaViT (Meng et al., 2022, adaptive depth-width-head selection), and Flash Attention V2 (Dao, 2023) provides exact attention with linear memory.

The emergence of Vision Transformers as the dominant computer-vision architecture is inseparable from the scaling laws they obey. Empirical scaling — increasing parameters, data, and compute together — has produced consistent improvements in ViT downstream performance up to the largest scale tested (21.7B parameters in ViT-22B). This section reviews what is known about ViT scaling, the systems engineering that makes it possible, and the frontier models of 2023–2026.

10.1. Empirical scaling: ViT-G, ViT-22B, EVA, DINOv2

Zhai et al. (NeurIPS 2022, "Scaling Vision Transformers") performed the most comprehensive ViT scaling study. They trained 50+ models with parameters spanning 5M (ViT-Ti) to 1.84B (ViT-G/14) and pretraining data spanning 30M to 3B images on JFT-3B. Their headline finding is that ImageNet error decreases as a power law in compute: $E \propto C^{-\alpha}$ with $\alpha \approx 0.16$ across roughly five orders of magnitude. The corresponding optimal model size grows as $N \propto C^{0.5}$ and the optimal training data as $D \propto C^{0.5}$ — exactly the same Chinchilla-style joint scaling that holds for language. ViT-G/14 with 1.84B parameters reaches 90.45% on ImageNet-1k (with extra fine-tuning at 518^2), 88.5% linear probe.

ViT-22B (Dehghani et al., ICML 2023) extends this scaling by an order of magnitude. The architectural changes from ViT-G are modest: parallel attention/MLP layers (rather than serial) for higher TPU throughput, query/key normalization to stabilize attention at width 6144 (without it, training diverges within 100 steps), omitted bias terms in QKV projections, and asynchronous data parallelism via Pathways. Training used 21,743M parameters, 17.4B JFT-3B examples seen, $\approx$ 170,000 TPU v4 core-days, 2.4M batch size, and the result was 89.51% ImageNet-1k linear-probe accuracy (in the same protocol where ViT-G reached 88.5%). Crucially, the loss curve of ViT-22B did not plateau, suggesting that further scaling would continue to deliver gains.

EVA-02 (Sun et al., 2023) reaches 90.0% ImageNet-1k top-1 with only 304M parameters by combining MIM pretraining with CLIP-distilled feature reconstruction targets and aggressive dropout. EVA-CLIP scales the dual-tower CLIP architecture to 5B parameters and reaches 79.0% IN-1k zero-shot, using only LAION-2B. DINOv2 (Oquab et al., 2023) trains ViT-S/14 (22M), ViT-B/14 (86M), ViT-L/14 (304M), and ViT-g/14 (1.1B) on the curated LVD-142M corpus; the ViT-g/14 model reaches 86.5% IN-1k linear probe and dominates dense-prediction transfer at frozen-feature evaluation.

For multimodal scaling, PaLI-X (Chen et al., 2023) combines a ViT-22B image encoder with a 32B mT5 language model into a 55B-parameter joint model. The resulting system reaches 88.5% on VQAv2, sets state-of-the-art on 25+ multilingual vision-language benchmarks, and substantially outperforms PaLI-3B (5B total) trained on the same data.

10.2. Mixture-of-experts and conditional computation

Sparse mixture-of-experts (MoE) is the principal mechanism by which the ViT field has begun to scale beyond ViT-22B without proportional compute increase. V-MoE (Riquelme et al., NeurIPS 2021) replaces selected MLP layers in a ViT with mixtures of E

experts, routing each token to the top-k experts ($k = 2$ typically). V-MoE-15B activates only 14.7B parameters per token while having 14.7B/0.5 = ≈ 29B total parameters; it matches a dense ViT-G with 25% less compute and reaches 90.35% on ImageNet-1k transfer. LIMoE adds language tokens to the same MoE router, demonstrating that vision and language can share experts.

Soft MoE (Puigcerver et al., ICLR 2024) replaces the top-k hard router with a soft assignment that is differentiable end-to-end, simplifying training and improving stability. Soft MoE ViT-H matches a dense ViT-22B on six VTAB tasks at one-fifth the compute. Sparse MoE has not yet achieved the dominance in vision that it has in language (where Mixtral, GLaM, and Switch Transformer drive frontier performance), partly because vision tokens are more uniform than language tokens and benefit less from routing, and partly because hardware-friendly inference of sparse models is an open problem.

Token-conditional computation, including DynamicViT, Evo-ViT, and AdaViT, provides a complementary form of conditional compute by skipping unimportant tokens at inference. Combined with weight-sparsity, these approaches deliver 50–80% inference FLOP reductions at <1% accuracy loss.

10.3. Training cost, energy, and memory budgets

Frontier ViT training has become computationally extraordinary. Approximate compute budgets (per published numbers and reasonable estimates):

These numbers expose two structural realities. First, vision Transformer training is now dominated by data ingestion and not by FLOPs per image: ViT-22B and PaLI-X reach the limits of available high-quality data more than they reach the limits of TPU pods. Second, the energy cost of training a single frontier ViT is comparable to the annual electricity consumption of 100–300 U.S. households, motivating efficiency research.

Memory is equally important. Training ViT-L at 1024×1024 resolution requires ≈ 40 GB of activation memory per GPU at batch size 4, exhausting an A100. Solutions include gradient checkpointing (4× memory reduction at 30% time cost), Flash Attention (Dao et al., 2022 — exact attention with $\mathcal{O}(N)$ memory; Flash Window Attention by Zhang 2025 specializes this for Swin), bf16 mixed precision, and ZeRO sharding via DeepSpeed/FSDP. Inference of ViT-22B fits on a single TPU v4 chip only in 8-bit quantization with FlashAttention.

10.4. What scaling laws teach us about ViT design

The scaling-laws literature has produced several robust empirical claims about Vision Transformers that practitioners should treat as defaults.

(i) ViT scaling is well-approximated by power laws over 5+ orders of magnitude in compute, with exponent $\alpha \approx 0.16$ for classification error and $\alpha \approx 0.20$ for downstream transfer error.

(ii) Plain ViTs scale better than hierarchical ViTs above 1B parameters: Swin V2 plateaued around 3B parameters, while plain ViT-22B continues to improve.

(iii) Patch size matters less than total tokens: ViT-G/14 outperforms ViT-G/16 by 0.3% IN-1k at the same FLOPs, consistent with finer tokenization at high resolution.

(iv) MAE pretraining benefits from longer training but with diminishing returns: 1600 epochs is the sweet spot for ViT-L on ImageNet-1k.

(v) Below 100M parameters, hybrids (CoAtNet, MobileViT, Mobile-Former) consistently win because the inductive bias gain dominates the slightly worse scaling.

(vi) Combining objectives (CLIP + MIM in EVA-02; SigLIP + MIM in OpenCLIPv2) yields a 1–2% boost over each in isolation, suggesting that pretraining objectives carry complementary information.

The main open scientific question is whether scaling will continue to pay off at the next order of magnitude (200B parameters, trillions of images). PaLI-X is the only published frontier multimodal model approaching that scale. Section 11 next interrogates the robustness and failure modes of these very large models.

11. Robustness, Safety, and Failure Modes of Vision Transformers

Whereas Section 10 focused on positive scaling, this section turns to failure modes across four parts: adversarial robustness and patch attacks, out-of-distribution generalization, calibration and hallucination, and ViT-specific safety risks. Adversarial work includes Shao et al. (2021, PGD on ViT-B/16), Bai et al. (2021, "Are Transformers More Robust Than CNNs?"), Paul and Chen (2022, "Vision Transformers Are Robust Learners" on IN-A/R/Sketch), Naseer et al. (2021, shape-bias and

Model	Params	Pretraining tokens (images)	Compute	Energy
ViT-B/16 (IN-21k)	86M	14M × 90 epochs ≈ 1.3B	≈ 100 TPU-days (v3)	≈ 1 MWh
ViT-L/16 (JFT-300M)	307M	300M × 7 epochs ≈ 2.1B	≈ 680 TPU-days	≈ 7 MWh
ViT-H/14 (JFT-300M)	632M	300M × 14 epochs ≈ 4.2B	≈ 2.5k TPU-days	≈ 25 MWh
MAE ViT-H (IN-1k)	632M	1.28M × 1600 epochs ≈ 2B	≈ 1.6k V100-days	≈ 17 MWh
CLIP ViT-L/14	304M (img)	400M × 32 epochs ≈ 12.8B	≈ 1.5k V100-days	≈ 16 MWh
ViT-G/14 (JFT-3B)	1.84B	3B × 5 epochs ≈ 15B	≈ 16k TPU-days	≈ 170 MWh
ViT-22B	21.7B	17.4B examples	≈ 170k TPU v4-days	≈ 1800 MWh
PaLI-X 55B	55B	mixed	≈ 250k+ TPU v4-days	≈ 3000 MWh
SAM ViT-H (data+train)	636M	11M images × 90 epochs	≈ 256 A100-days	≈ 3 MWh
DINOv2 ViT-g/14	1.1B	142M × ~10 epochs	≈ 22k A100-hrs	≈ 7 MWh
Stable Diffusion 3 (DiT)	≈ 8B	≈ 1B+	(proprietary)	≈ tens of MWh

patch attacks), SmoothViT (2022, randomized patch ablation), Salman et al. (2022, certified patch defense), Zhao et al. (2022, token-level certified via masked-patch ablation), and Chen and Liu (2022, Holistic Adversarial Robustness catalog). Attack-SAM (Zhang et al., 2023, prompt and pixel attacks on SAM), UniGuardian (Lin et al., 2025, multimodal LVLM attack catalog), and Darcet et al. (2024, "ViTs Need Registers" attention-sink fix) widen the failure-mode catalog. Calibration work includes Minderer et al. (2021, ViT ECE) and Park et al. (2026, DETR object-level calibration); CLIP-Guided Decoding (Deng et al., 2024, hallucination reranking), DN-DETR (Li et al., 2022, denoising bipartite matching), and Co-DETR (Zong et al., 2024, one-to-many auxiliary matching) address generation and DETR failure modes. Multimodal robustness is addressed by DisCo-CLIP (Chen et al., 2023, contrastive backdoor defense), Tu et al. (2024, holistic CLIP robustness over 17 shifts), You et al. (2024, group-robust multi-modal calibration), and Nakajima et al. (2025, neuroscience-inspired adversarial defense).

Vision Transformers are now deployed in safety-critical pipelines including autonomous driving, medical imaging, biometric authentication, and content moderation. Their failure modes are therefore under increasing scrutiny. This section reviews the empirical record on adversarial robustness, distribution-shift robustness, and calibration. It also covers the qualitatively new failure modes introduced by Transformer-specific design choices such as patch tokenization and DETR-style set prediction.

11.1. Adversarial robustness and patch attacks

Whether ViTs are intrinsically more robust than CNNs to adversarial perturbations was a major question in 2021 and remains partly open. The first systematic study by Shao et al. (2021) showed that under untargeted PGD attack with ℓ_∞ budget $\varepsilon = 4/255$ and 10 iteration steps, ViT-B/16 dropped from 84.0% clean accuracy to 0.05% adversarial accuracy on ImageNet-1k validation, similar to ResNet-50 (which dropped from 76.1% to 0.04%). Bai et al. (2021), in "Are Transformers More Robust Than CNNs?", reported that the modest natural-corruption advantage of ViTs over CNNs largely vanishes under standardized adversarial training: adversarially trained DeiT-B and ResNet-50 reach within 1.5 percentage points of each other on AutoAttack at $\varepsilon = 4/255$.

However, Paul and Chen (AAAI 2022, "Vision Transformers Are Robust Learners") demonstrated a clear ViT advantage on natural-corruption benchmarks. Without adversarial training, ViT-L/16 outperforms ResNet-152 by 8 points on ImageNet-A, 35 points on ImageNet-R, 24 points on ImageNet-Sketch, and reduces mean corruption error on ImageNet-C from 76.7 (ResNet-50) to 47.0 (ViT-L). Subsequent work attributed this gap to two factors: (i) ViT's global receptive field aggregates information across the whole image, suppressing the influence of localized texture cues that CNNs over-rely on; and (ii) ViT's lower dependence on shape-vs-texture biases — the original Geirhos et al. shape-bias study (later extended to ViT by Naseer et al., NeurIPS 2021) found that ViT-B/16 has a shape-bias of 28% versus ResNet-50's 17%.

A specifically Transformer-relevant attack family is

patch attacks, in which an adversary perturbs only one or a few patches rather than the full image. Because ViT's tokenization treats each patch as a unit, a single perturbed patch can alter all attention scores. Naseer et al. and Fu et al. showed that adversarial patches at training resolution drop ViT-B/16 from 84% to <30% accuracy with attacks affecting only 4% of pixels — substantially worse than ResNet under equivalent area constraints. Defenses include patch-based smoothing (Salman et al., NeurIPS 2022), randomized patch ablation, and SmoothViT.

Token-level certified robustness has been studied by Zhao et al., who showed that a tight ℓ_∞ certification can be obtained by averaging predictions over masked-patch ablations, yielding 26% certified robust accuracy at $\varepsilon = 0.1$ on ImageNet for ViT-B/16. Holistic Adversarial Robustness (Chen and Liu, 2022) catalogs more than 30 attack-defense pairs across CNNs, ViTs, and language-vision models.

11.2. Out-of-distribution generalization on ImageNet-A/R/Sketch/C

The OOD generalization story for ViTs is nuanced. CLIP-pretrained ViTs are dramatically more robust than ImageNet-supervised ViTs: CLIP ViT-L/14 scores 70.7% on ImageNet-A and 87.1% on ImageNet-R zero-shot, vs. supervised ViT-L at 28% and 49% respectively. The robustness premium of CLIP is attributable to (i) its diverse 400M-image pretraining corpus, (ii) the open-vocabulary supervision provided by natural-language captions, and (iii) the absence of explicit ImageNet category boundaries that supervised models overfit. Tu et al. (2024) holistically evaluated CLIP robustness across 17 distribution shifts and found a 25-percentage-point average advantage over supervised ViT-L.

DINOv2 features, when used with a linear probe trained on ImageNet, show OOD numbers comparable to CLIP: 56.7% on ImageNet-A, 80.7% on ImageNet-R, suggesting that diverse self-supervised pretraining alone confers substantial OOD robustness even without language supervision.

ViTs, however, exhibit a known failure mode that CNNs do not: attention sinks and degenerate query distributions. Darcet et al. (ICLR 2024, "Vision Transformers Need Registers") observed that pretrained ViTs frequently allocate disproportionate attention mass to a small number of background patches that contain no semantic information; these "register tokens" emerge during training as a side-channel for global computation. Adding explicit learnable register tokens to the input sequence eliminates the artifacts and improves attention-map quality for downstream segmentation and depth estimation by 3–8% on dense-feature benchmarks.

11.3. Calibration, hallucination, and shortcut learning

Calibration — the degree to which a model's predicted probabilities reflect its true accuracy — is essential for medical and safety-critical use. Empirically, ViTs are slightly better calibrated than CNNs: Minderer et al. (2021) report Expected Calibration Error (ECE) of 0.018 for ViT-L/16 vs. 0.044 for ResNet-152 on ImageNet, but both still benefit from temperature scaling. Detection Transformer calibration was studied by Park et al. (TPAMI 2026), who showed that DETR-family detectors generate hundreds of low-confidence proposals per image, and proposed object-level calibration plus image-level reliability scoring to mitigate.

Vision-language Transformers introduce a qualitatively new failure mode: object hallucination in image captioning. Models like BLIP-2, LLaVA, and InternVL generate text that mentions objects not present in the image, with reported hallucination rates of 5–15% on standard benchmarks. CLIP-Guided Decoding (Deng et al., 2024) reduces hallucination by reranking generated tokens against CLIP image-text similarity, cutting hallucination rate by 40% on POPE and CHAIR benchmarks. Object-guided CLIP (Zheng, 2026) and similar work tackle the converse problem of CLIP missing salient objects in cluttered scenes.

DETR-family detectors share a related failure mode known as the dual assignment problem: bipartite matching can be unstable, especially in early training, leading to redundant predictions and missed detections of partially occluded objects. DN-DETR and DINO-DETR address this with denoising training; Co-DETR uses one-to-many matching during training to stabilize gradients while preserving the one-to-one matching that gives end-to-end behavior.

A pattern of shortcut learning affects all large vision models. ViT-22B's unprecedented ImageNet accuracy was accompanied by surprisingly weak performance on fine-grained bird species (CUB-200-2011), suggesting that some of the gain at scale comes from learning short-cut features rather than canonical object-shape priors. In medical imaging, Goh et al. (2024) compared ViTs and CNNs on referable diabetic retinopathy and found ViTs over-relied on dataset-specific photographic artifacts (vignetting, illumination gradients), missing borderline disease cases.

Failure mode	Affected models	Mitigation	Citation
Adversarial PGD attack	All ViTs	Adversarial training	Bai et al. 2021
Patch adversarial attack	ViT (tokenization)	SmoothViT, ablation defense	Naseer et al. 2021
Attention-sink registers	Plain ViTs	Register tokens	Darcet et al. 2024
Distribution shift IN-A	Supervised ViT	CLIP / DINOv2 pretraining	Hendrycks 2021
Calibration drift	All ViTs	Temperature scaling	Minderer 2021
DETR redundant predictions	DETR family	DN-DETR, Co-DETR	Zhang 2023, Zong 2024
Object hallucination	LVLMs (BLIP-2, LLaVA)	CLIP-guided decoding	Deng 2024
Shortcut learning	Foundation ViTs	Domain-specific FT	Goh 2024
Backdoor / poisoning	CLIP	DisCo-CLIP, robust train	Chen 2023
Patch-attack on detection	DETR	Cert. patch defense	Salman 2022

11.4. Patches that pose unique safety risks

Beyond accuracy-degrading attacks, ViT-specific attacks have been demonstrated against safety-critical applications. UniGuardian (Lin et al., 2025) catalogs prompt injection, backdoor, and adversarial attacks on multimodal LVLMs that would not apply to a pure vision system. CLIP-TD and similar distillation works show that adversarial CLIP examples transfer to downstream segmentation systems built on top of CLIP. Attack-SAM (Zhang et al., 2023) demonstrates that crafted prompts and pixel perturbations can entirely block SAM's segmentation output, raising concerns for safety-critical reliance on the SA-1B-trained ViT-H encoder.

Safety mitigations are now a mainstream research direction. Adversarial defense via brain-activity integration (Nakajima et al., 2025) explores neuroscience-inspired robustness. Calibrating multi-modal representations (You et al., 2024) addresses group robustness without requiring explicit group labels. Toward a Holistic Evaluation of Robustness in CLIP Models (Tu et al., 2024) proposes a 7-axis evaluation suite covering visual, textual, and modality-bridging perturbations.

The summary picture is that Vision Transformers are not categorically more or less robust than CNNs; they exhibit a different robustness profile. They tend to be more robust to natural distribution shifts and more vulnerable to patch-localized attacks, more uniformly calibrated but more prone to attention-sink artifacts, and more powerful in the multimodal regime but susceptible to language-driven hallucinations. Practitioners should choose pretraining objectives and fine-tuning protocols with these failure modes in mind. Section 12 turns to concrete application domains where these trade-offs play out.

12. Application Domains: Medical Imaging, Remote Sensing, Autonomous Driving, Robotics

Building on Section 11, this section turns to deployed application clusters across four domain groups: medical imaging, Earth observation, autonomous driving and robotics, and industrial applications. Medical imaging includes Swin UNETR (Hatamizadeh et al., 2022, 3D Swin), Swin-Unet (Cao et al., 2023, 2D pure-Transformer U-Net), TransUNet (Chen et al., 2021, hybrid CNN-Transformer), TransDeepLab (Azad et al., 2022, Swin DeepLab), Swin-UMamba (Liu et al., 2024, Mamba medical), and CIS-UNet (Imran et al., 2024, contextual UNet). The SAM-derived lineage includes MedSAM (Ma et al., 2024), SAM-Med2D and SAM-Med3D (Wang et al., 2025), MedicoSAM (Archit et al., 2025), and SurgiSAM2 (Kamtam et al., 2025) for surgical video. Foundation-model deployments include RETFound (Zhou et al., 2023, retina), UNI, CONCH, and Virchow (Vorontsov et al., 2024, pathology), Swin-GA-RF (Alohali et al., 2024, cervical cancer), and Mi-M (Zhuang et al., 2025, 3D tokenization). Earth observation includes Prithvi-100M (NASA-IBM, 2024, MAE geospatial), HyperSIGMA (Wang et al., 2025, hyperspectral), RingMo (Sun et al., 2023, Chinese RS), DOFA-CLIP (Xiong et al., 2025, multimodal Earth observation), and Aquatic-CLIP (Alawode et al., 2026, underwater); SwinSight (Pradhan et al., 2024) targets building extraction, SwinUNet3D (Bojesomo et al., 2022) handles traffic prediction, and Siamese Swin-Unet (Tang et al., 2024) addresses change detection. Autonomous driving includes BEVFormer and BEVFormer v2 (Li et al., 2022/2023), PETR and PETR v2 (Liu et al., 2022, 3D-position queries), Sparse4D (Lin et al., 2023, 4D query attention), V2X-ViT (Xu et al., 2022, cooperative perception), SKGE-Swin (Kartiman et al., 2025), UniAD

(Hu et al., 2023, end-to-end planning), VAD (Jiang et al., 2023, vectorized driving), and DriveGPT-4 (Xu et al., 2024, driving LLM). Robot learning includes the MAE-pretrained ViT for robots (Radosavovic et al., 2022), VLS (Liu et al., 2026, VLM-steered diffusion), RT-2 (Brohan et al., 2023), and PaLM-E (Driess et al., 2023, embodied multimodal). Industrial and biometric applications include ConViTX (Karthik et al., 2024, plant disease), HC-ViT (Zhang et al., 2024, transmission-line defect), SAID (Huang et al., 2025, industrial defect), ViT-FIQA (Atzori et al., 2025, face quality), and MAE-DFER (Sun et al., 2023, dynamic facial expression).

Vision Transformers have been adopted across virtually every domain that previously used CNNs. This section reviews four illustrative application clusters — medical imaging, Earth observation, autonomous driving, and robotics — that account for the bulk of recent ViT-derived work and that demonstrate how the architecture choices reviewed in earlier sections translate into deployed systems.

12.1. Medical imaging foundation models: Swin UNETR, MedSAM, RETFound, UNI

Medical imaging is the largest application area by paper count, with over 90% of medical-image data being two-dimensional (X-rays, fundus images, dermoscopy) or three-dimensional volumetric scans (CT, MRI). Al-Hammuri et al. (2023) and He et al. (Intelligent Medicine 2022) provide comprehensive surveys. We highlight five archetypal systems.

Swin UNETR (Hatamizadeh et al., MICCAI 2022) combines a Swin Transformer encoder with a U-Net-style convolutional decoder for 3D volumetric segmentation. It won the BraTS 2021 challenge for brain tumor segmentation, achieving Dice scores above 0.9 on whole-tumor and 0.85 on tumor-core regions. Swin UNETR has since been applied to head-and-neck cancer, abdominal organ segmentation (AMOS22), and whole-body PET, with consistent gains over the nnU-Net CNN baseline.

Swin-Unet (Cao et al., ECCV Workshops 2023) is a pure-Transformer 2D U-Net analogue with shifted-window attention, reaching state-of-the-art on the Synapse multi-organ benchmark with mean Dice 0.79 and on ACDC cardiac segmentation with 0.93. TransUNet (Chen et al., 2021), TransDeepLab (Azad et al., 2022), Swin-UMamba (Liu et al., 2024), and CIS-UNet (Imran et al., 2024) are derivative architectures.

MedSAM (Ma et al., Nature Communications 2024) fine-tunes SAM on 1.57M medical image-mask pairs spanning 10 imaging modalities and 30 cancer types, producing the first medical foundation segmentation model. MedSAM achieves Dice scores within 5% of task-specific baselines on 86 internal tasks while being a single zero-shot model. SAM-Med2D, SAM-Med3D (Wang et al., 2025), MedicoSAM (Archit et al., 2025), and the SAM 2-based SurgiSAM2 (Kamtam et al., 2025) extend this lineage. Practical evaluations such as Huang et al. (Medical Image Analysis 2023) show that out-of-the-box SAM struggles in 3D medical imaging without prompts, and that fine-tuning is generally required.

RETFound (Zhou et al., Nature 2023) is a foundation model for retinal images, pretrained with MAE on 1.6M unlabeled fundus and OCT images and fine-tuned for diabetic retinopathy, glaucoma, age-related macular degeneration, and Parkinson's disease detection. RETFound exceeds the AUC of supervised baselines by 5–15 points across nine tasks while requiring 10× less labeled fine-tuning data.

UNI, CONCH, and Virchow (Vorontsov et al., Nature Medicine 2024) are foundation models for computational pathology, all built on ViT-L/14 or ViT-g/14 backbones pretrained with DINOv2-style self-distillation on 100M+ histopathology tile images. UNI was trained on 100M tiles from 100,000 H&E whole-slide images and reaches AUC 0.95+ for predicting microsatellite instability and 0.91 for HER2 status — surpassing rapidly any task-specific CNN trained on the same data. Subsequent fine-grained applications include cervical cancer classification (Swin-GA-RF; Alohali et al., 2024), thyroid nodule detection, and oral disease detection.

Domain-specific challenges for medical ViTs include 3D volumetric tokenization (Mi-M; Zhuang et al., 2025), arterial-spin-labeling MRI denoising (Cheema et al., 2025), and motion artifact correction (Hossain et al., 2024). The dominant trend in 2025–2026 is foundation models followed by parameter-efficient adaptation via LoRA, adapters, or prompt tuning.

12.2. Earth observation: Prithvi, HyperSIGMA, RingMo, ViT-RS

Aleissaee et al. (Remote Sensing 2023) survey 80+ ViT-based remote-sensing systems. Bazi et al. (Remote Sensing 2021) introduced ViT to remote sensing image classification, achieving 98.4% on the AID benchmark.

Prithvi-100M is a 100M-parameter NASA-IBM geospatial foundation model pretrained with MAE on 4.2 billion HLS (Harmonized Landsat-Sentinel) pix-

els at 30 m resolution, demonstrated for flood mapping, wildfire scar detection, and crop classification. HyperSIGMA (Wang et al., TPAMI 2025) is a hyperspectral foundation model with billions of pixels of training data across hundreds of spectral channels, and RingMo is a Chinese remote-sensing foundation model with multi-spectral pretraining on 2M satellite tiles. DOFA-CLIP (Xiong et al., 2025) extends CLIP to multimodal Earth-observation, aligning optical, radar, multispectral, and hyperspectral data with text, and AquaticCLIP (Alawode et al., 2026) targets underwater imagery.

For tasks like building extraction, the multiscale-feature SegFormer adaptations (Chen et al., 2022) and SwinSight (Pradhan et al., 2024) lead in mIoU on AIRS and Massachusetts Buildings. SwinUNet3D (Bojesomo et al., 2022) and Siamese Swin-Unet (Tang et al., 2024) handle traffic prediction and change detection respectively. Bolcek et al. (2024) provide a comprehensive evaluation of deep ViTs for road extraction from very-high-resolution satellite imagery, finding ViT-Adapter-L the strongest single model.

12.3. Autonomous driving and robotics: V2X-ViT, BEVFormer, robot-learning ViTs

Autonomous driving deploys ViTs at multiple stages of the perception pipeline. BEVFormer (Li et al., ECCV 2022) and BEVFormer v2 project multi-camera images to a bird's-eye-view feature grid through deformable cross-attention; on nuScenes test, BEVFormer reaches 51.7 NDS, beating LiDAR baselines like CenterPoint at 50.4 NDS. DETR3D lifts DETR's 2D queries to 3D space, reaching 41.7 NDS, while PETR and PETR v2 embed 3D positional encodings into 2D image features and Sparse4D uses 4D query attention with temporal aggregation to reach 53.6 NDS.

V2X-ViT (Xu et al., ECCV 2022) tackles the cooperative-perception problem in vehicle-to-everything communication. The system fuses ego-vehicle features with features broadcast by neighboring vehicles and roadside units through a Transformer attention module, robust to communication delay and pose error. V2X-ViT improves IoU by 21% over single-vehicle baselines on the V2XSet simulation benchmark.

SKGE-Swin (Kartiman et al., 2025) applies Swin Transformer to end-to-end waypoint prediction for autonomous-driving navigation, with skip-stage connections to retain global context. End-to-end driving Transformers like UniAD, VAD, and DriveGPT-4 fuse perception, prediction, and planning in a single multimodal Transformer, with UniAD reaching 16.4% lower planning error than modular baselines on nuScenes.

For robot learning, Real-World Robot Learning with Masked Visual Pre-training (Radosavovic et al., 2022) demonstrated that MAE-pretrained ViTs outperform supervised ImageNet ViTs as visual encoders for robotic manipulation, with 30% higher task success on real-world reaching, grasping, and pushing benchmarks. VLS (Liu et al., 2026) uses pretrained vision-language models to steer robot diffusion policies near obstacles, and RT-2 and PaLM-E from Google integrate ViT-22B-derived encoders with language and action heads for general-purpose robot control.

12.4. Other application clusters

In agriculture, ConViTX targets plant disease identification, SwinConvNeXt addresses garbage classification, Karthik et al. (2024) study grape leaf disease, and Chen et al. (2026) tackle grain counting. Surveillance and biometric work includes ViT-FIQA for face image quality (Atzori et al., 2025) and human action recognition surveys by Alomar et al. (2025). Industrial inspection is represented by HC-ViT for transmission-line defect detection (Zhang et al., 2024) and SAID for industrial defect segmentation (Huang et al., 2025). Pathology and microscopy applications include Vim4Path (Nasiri-Sarvi et al., 2024), SegmentCellSAM, Hi-End-MAE (Tang et al., 2025), and the all-cell-in-SAM model. For texture analysis, the comparative ViT survey of Scabini et al. (Journal of Imaging 2025) found ViT-L/14 to dominate classical Gabor-feature CNNs by 12% on KTH-TIPS. Music and time-series work is exemplified by MTSMAE (Tang and Zhang, 2022), and affect computing draws on Action Transformer (Mazzia et al., 2021) and MAE-DFER (Sun et al., 2023).

These applications make a uniform engineering point: the ViT recipe transfers with little modification across domains as long as a sufficiently large pretraining corpus exists. Where domain-specific data is scarce, the dominant pattern is to start from a frozen DINOv2, CLIP, or SAM image encoder and add a lightweight adapter or fine-tune a parameter-efficient subset (LoRA, prompt tuning, BitFit). The success of MedSAM, RETFound, UNI, and Prithvi all follow this template. Section 13 closes the survey with our predictions about what will replace or extend these systems.

Domain	Representative ViT system	Backbone	Task	Dataset	Headline metric
Medical 3D segmentation	Swin UNETR	Swin-B 3D	Brain tumor seg	BraTS 2021	0.91 Dice WT
Medical 2D segmentation	Swin-Unet	Swin-T	Multi-organ	Synapse	0.79 Dice
Medical foundation seg	MedSAM	ViT-B / SAM	30 cancer types	1.57M pairs	within 5% of specialist
Retinal foundation	RETFound	ViT-L MAE	DR, AMD, glaucoma	1.6M images	AUC > 0.95
Pathology foundation	UNI	ViT-L DINOv2	MSI, HER2	100M tiles	AUC 0.95+
Remote sensing	Prithvi-100M	ViT MAE	Flood, wildfire	4.2B HLS pixels	n/a (zero-shot transfer)
Hyperspectral	HyperSIGMA	ViT	Land cover	100s of channels	TPAMI 2025
Autonomous driving	BEVFormer	ResNet-101 + Tx	3D detection	nuScenes	51.7 NDS
V2X cooperative	V2X-ViT	PointPillar + Tx	Cooperative perc.	V2XSet	+21% IoU
Robot manipulation	MAE-pretrained ViT	ViT-B	Reach/grasp/push	Real robot	+30% success
Plant disease ID	ConViTX	ViT + CNN	Leaf classification	PlantVillage	99.0% acc
Underwater	AquaticCLIP	CLIP ViT	Underwater scenes	TNNLS 2026	n/a
Face image quality	ViT-FIQA	ViT	FIQA	n/a	SOTA
Industrial defect	SAID	SAM-derived	Surface defects	n/a	n/a
Surgical video	SurgiSAM2	Hiera (SAM 2)	Anatomy seg	Multi-center	n/a

13. Open Problems and Falsifiable Predictions for 2026 and Beyond

Whereas Section 12 surveyed deployed applications, this section turns to open problems and falsifiable predictions across four parts: alternatives to softmax attention, tokenization futures, evaluation, and a structured table. The state-space frontier includes Mamba (Gu and Dao, 2023, selective state-space scan), Vision Mamba or Vim (Zhu et al., 2024, bidirectional Mamba on patches), VMamba (Liu et al., 2024, two-direction visual Mamba), Swin-UMamba (Liu et al., 2024, Mamba U-Net medical), Vim4Path (Nasiri-Sarvi et al., 2024, Mamba pathology), Hi-End-MAE (Tang et al., 2025, hierarchical MAE-Mamba), and Kolmogorov-Arnold Transformers or KAT (Yang and Wang, 2024, KAN MLP replacement). Tokenization futures include NaViT (Dehghani et al., 2023, native-aspect packing) and ViT-5 (Wang et al., 2026, modernized canonical recipe). Conditional-compute work includes Soft MoE (Puigcerver et al., 2024, differentiable expert routing), V-MoE (Riquelme et al., 2021, top-k vision MoE), and Register-Token ViT (Darcet et al., 2024, attention-sink fix). Systems-level advances include Flash Attention V2 (Dao, 2023, exact attention with linear memory), Flash Window Attention (Zhang, 2025, fused-kernel Swin), and Object-guided CLIP (Zheng, 2026, salient-object retrieval grounding).

The Vision Transformer field at the time of writing (mid-2026) is mature enough that incremental architectural changes deliver diminishing returns, yet several large open problems remain. This section closes the survey with a structured set of forecasts. Each prediction is intended to be falsifiable: a specific benchmark, model, or system whose appearance (or non-appearance) by a stated horizon will confirm or refute the claim. We also provide a glossary of terms used throughout the survey.

13.1. Beyond softmax: state-space, linear, and Mamba alternatives

The clearest near-term threat to the Transformer's dominance is the rise of selective state-space models. Mamba (Gu and Dao, 2023) introduced a hardware-aware selective scan that achieves $\mathcal{O}(N)$ compute and memory while approximating attention's expressivity. Vision Mamba (Vim; Zhu et al., 2024) and VMamba apply the operator to images by linearizing patch grids in two diagonal directions, reaching Swin-comparable accuracy at substantially better long-sequence efficiency. Vision Mamba-S reaches 80.5% IN-1k top-1 with 26M parameters, matching Swin-T at 28M; on ADE20K segmentation, Vim-S reaches 47.3 mIoU, within 0.3 mIoU of Swin-S at 47% the FLOPs. Swin-UMamba (Liu et al., 2024) and Hi-End-MAE (Tang et al., 2025) integrate Mamba with U-Net structures for medical segmentation. Vim4Path (Nasiri-Sarvi et al., 2024) outperforms ViT-UNI on whole-slide histopathology.

We predict (with falsification window 2027) that for tasks involving inputs longer than 4,000 tokens — high-resolution segmentation at 4K, hour-long video understanding, multi-document visual reasoning — Mamba-based architectures will displace Transformers in published SOTA. Hybrid Mamba-Transformer architectures (one Mamba block per Transformer block) are likely to dominate the medium-resolution regime where neither approach has a clear winner.

A second alternative is Kolmogorov-Arnold Transformers (Yang and Wang, 2024), which replace MLP layers with KANs (Kolmogorov-Arnold Networks). On ImageNet-1k, KAT-B reaches 83.6%, comparable to ViT-B. Whether the additional expressivity of learnable activation functions outweighs the higher compute cost remains unsettled.

13.2. Tokenization, NaViT, and adaptive resolution

Patch tokenization remains the most under-explored design choice in Vision Transformers. Fixed 16×16 patches were chosen by Dosovitskiy et al. for computational convenience and have been retained largely by inertia. Several recent works challenge this choice. NaViT (Dehghani et al., 2023) packs multiple images of varying aspect ratios into one sequence with attention masking; it accepts native-resolution inputs, eliminating the need for the standard 224^2 resize. NaViT-L reaches 88.5% on IN-1k while saving $5\times$ compute through aspect-ratio-aware packing. ViT-5 (Wang et al., 2026) unifies a series of post-2021 patch-stem improvements — overlapping patches, conv stems, learnable patch positions — and reports SOTA at matched FLOP budget across IN-1k, COCO, and ADE20K.

We predict (falsification window 2027) that fixed 16×16 patching will be replaced as the default by content-adaptive tokenization in three out of five major vision-Transformer papers per major venue. The natural endpoint is a "perceiver-style" architecture in which the input is a continuous field and tokens are sampled adaptively, dropping the rigid grid entirely.

13.3. Foundation-model evaluation and reproducibility

The benchmark stack reviewed in Section 9 was designed for an era of single-model, single-task evaluation. The advent of foundation models like CLIP, DINOv2, SAM, ViT-22B, and PaLI-X breaks the assumption that a single number characterizes a model. A frozen DINOv2 may dominate dense prediction while a CLIP at the same parameter count dominates zero-shot recognition; ViT-22B excels at scaling while SAM dominates segmentation prompts. The community has begun to assemble multi-axis evaluation suites — VTAB-1k for transfer, MMBench for multimodal reasoning, OOD-CLIP for distribution shift, MedAGI for medical foundations — but no consensus exists on how to report them collectively.

We predict that by 2027, conferences will require ViT papers to report at least four protocols (linear probe, k-NN, fine-tune, zero-shot) on at least three pretraining-dataset axes (in-domain, out-of-domain, cross-modal), and that papers reporting only ImageNet top-1 fine-tuning will no longer be publishable at NeurIPS, ICLR, or CVPR.

A second open problem is reproducibility. Of the JFT-pretrained models (ViT, ViT-G, ViT-22B, CoAtNet-7), none have publicly released training code or weights; LAION-pretrained CLIP-style models are reproducible but have weaker downstream performance. We predict that the gap between publicly reproducible and proprietary frontier models will close to <2 IN-1k accuracy points by 2027 as LAION-style data efforts continue.

13.4. Open problems and bottlenecks

The principal scientific bottlenecks for the next phase of ViT research are:

13.5. Concluding synthesis

The Vision Transformer field has, in nine years, moved through three distinct phases. The 2017–2020 preparatory phase saw attention as an auxiliary

Open problem	Current status	Falsifiable prediction (2027)
Quadratic attention cost	Linear & state-space alternatives emerging	Mamba-based ViT wins ≥1 major benchmark
Patch tokenization rigidity	NaViT, ViT-5 explore alternatives	Content-adaptive tokenization standard
Position-encoding extrapolation	RoPE, ALiBi for ViT trialed	High-res transfer >2× pretrain res routine
Adversarial robustness	PGD-vulnerability persists	Certified $\varepsilon \geq 8/255$ reached on ViT
Patch attacks	4%-pixel attacks fool ViT	Patch-defense >70% acc on adversarial patches
Object hallucination in LVLMs	5-15% rate	<3% with retrieval-grounded decoding
Attention-sink registers	Solved by register tokens (Darcet 2024)	Register tokens become ViT default
Video memory compression	SAM 2 memory bank limits videos to ~1 min	Hour-long video segmentation routine
Frontier compute energy	ViT-22B used ≈170k TPU-days	50% energy reduction via 4-bit pretraining
Foundation-model evaluation	Fragmented benchmark stacks	Unified VTAB-2 / MMBench-2 standard
Open scientific reproducibility	JFT corpus closed	Open 5B+ image dataset matches JFT
Sparse mixture-of-experts	V-MoE, Soft MoE early	MoE ViT routine in production
3D and 4D foundation models	SAM 2 video, Prithvi geo	4D radar-LiDAR Transformer foundation
Calibration on long-tail	DETR redundant predictions	Calibrated set-prediction with ECE<0.02
Domain transfer (medical/EO)	Hand-tuned LoRA adapters	Universal modality adapter via SigLIP-style
Edge deployment <2W	EfficientFormer V2	ViT-equivalent at 0.5W power

mechanism on top of CNNs. The 2020–2022 foundational phase introduced patch tokenization, hierarchical Transformers, masked image modeling, and contrastive vision-language learning, establishing the architectural blueprint of modern computer vision. The 2022–2026 expansion phase produced foundation models at every scale — ViT-22B, DINOv2, SAM, SAM 2, PaLI-X — and reorganized downstream tasks around prompt-based interaction with these foundations.

Looking forward, we identify six central tensions that will shape the next half-decade:

1. Architecture vs. data: every accuracy gain to date has required both. The field is approaching the limit of publicly available image-text data, and the next gains will require new data sources (synthetic data from generative models, web video, embodied robot rollouts).

2. Plain vs. hierarchical: plain ViT plus adapter has nearly closed the gap with hierarchical ViT for dense prediction; we expect plain ViTs to dominate at scale and hierarchical ViTs to persist at the mobile edge.

3. Softmax attention vs. linear / state-space: this is the most active architectural question. Mamba-style models look likely to dominate long-sequence vision but may not displace softmax for moderate sequence lengths.

4. Generative vs. discriminative pretraining: MAE, DINOv2, and CLIP each excel on different downstream tasks. The unified objective — generative + discriminative + contrastive — is the empirical winner (CoCa, EVA-02), and we expect it to standardize.

5. Single-modal vs. multimodal foundation: pure-vision ViT pretraining is increasingly subsumed by image-text dual-tower training. The frontier of "vision Transformer" will be indistinguishable from the frontier of "vision-language model" by 2027.

6. Research reproducibility vs. industrial scale: the

gap between open-source LAION-pretrained models and proprietary JFT/WebLI-pretrained models is the principal threat to the field's scientific health. Closing it requires sustained open-source data efforts.

13.6. Glossary

The Vision Transformer is no longer a radical experiment but the load-bearing architecture of computer vision. Its theory and practice will continue to evolve, but the central pattern — patch tokenization plus self-attention plus foundation-scale pretraining — is now the default. We hope this survey serves as both a tutorial entry point and a reference manual for the next wave of researchers and practitioners building on this foundation.

14. Critical Synthesis

Building on the predictions in Section 13, this section delivers an explicit cross-family comparison and a consolidated list of open problems. It is organized as three parts: a comparative analysis of major method families, a list of open problems spanning 2025–2026, and a list of future directions emerging this year.

14.1. Comparison across method families

The four backbone families compared in Section 3 trade off different axes. Plain ViTs (ViT, DeiT, MAE, DINOv2, ViT-22B) maximize scaling efficiency and transfer cleanly to MIM and self-distillation, but are weak below 100M parameters. Hierarchical ViTs (Swin, PVT v2, MViT v2, NAT, CSWin) deliver the best dense-prediction accuracy at moderate scale, with linear cost in image size, but plateau above roughly 3B parameters. Convolution-attention hybrids (CvT, CoAtNet, MetaFormer, ViTAEv2) win consistently below 100M parameters because the locality prior offsets data scarcity. Mobile and edge ViTs (MobileViT, Mobile-Former, EfficientFormer V2, EfficientViT) trade peak accuracy for sub-2 ms inference latency. Across these methods, the empirical winner depends on the data and compute budget rather than any single architectural property.

The three pretraining paradigms compared in Section 5 also trade off cleanly. Supervised JFT-scale classification (ViT, ViT-G, ViT-22B) gives the highest fine-tuning accuracy on ImageNet at every parameter count, but produces features that are more class-boundary-bound than alternatives. Masked image modeling (MAE, SimMIM, BEiT, MaskFeat, ConvNeXt V2, VideoMAE) gives the best fine-tuning accuracy when only ImageNet-1k is available and recovers most of the JFT advantage at lower data cost. Self-distillation (DINO, iBOT, DINOv2) gives the best frozen-feature linear-probe and dense-prediction transfer, especially for tasks lacking labels. Contrastive image-text pretraining (CLIP, ALIGN, OpenCLIP, SigLIP, EVA-CLIP, CoCa) gives the best zero-shot recognition and is the only paradigm that produces an open-vocabulary recognizer. Across these methods, the strongest empirical recipes (EVA-02, CoCa, ViT-22B with sigmoid loss) combine two or more paradigms.

The DETR family in Section 7 also illustrates a clear trade-off. DETR pioneered set prediction but converged slowly at 500 epochs. Deformable DETR cut training to 50 epochs at the cost of a sparser receptive field. DN-DETR and DINO-DETR added denoising training to stabilize bipartite matching. Co-DETR raised peak accuracy to 66.0 box AP via auxiliary one-to-many matching. RT-DETR finally reached real-time inference at 114 FPS by hybrid encoders and IoU-aware query selection. Crucially, the DETR family has converged to a common template — denoising plus mixed query selection plus deformable cross-attention — that should be considered the modern detection default. The state-space alternative, Vision Mamba and VMamba, currently matches Swin at lower compute on ImageNet and ADE20K but has not yet displaced softmax attention on COCO detection or VQAv2.

14.2. Open problems in 2025–2026

The principal open problems for ViT research in 2025–2026 are:

- Quadratic attention cost remains prohibitive for 4096^2 inputs; selective state-space alternatives like Vision Mamba and VMamba close part of the gap but do not yet dominate on COCO or VQAv2.
- Patch tokenization rigidity — fixed 16×16 patches inherited from 2020 — leaves resolution and aspect-ratio efficiency on the table; NaViT and ViT-5 offer partial fixes.
- Adversarial robustness against patch-localized attacks remains unsolved; certified defenses cap at $\varepsilon \leq 4/255$ at present.
- Object hallucination in vision-language models like BLIP-2, LLaVA, and InternVL persists at 5–15% and lacks a deployable decoding-time fix beyond reranking.
- Long video understanding beyond approximately one minute is bottlenecked by SAM 2's memory

bank; Mamba-style temporal compression is the leading candidate.

- Frontier compute energy — about 170,000 TPU v4-days for ViT-22B — is environmentally unsustainable; 4-bit pretraining and sparse MoE are the principal cost-reduction directions.
- Foundation-model evaluation is fragmented across VTAB, MMBench, OOD-CLIP, and MedAGI; a unified multi-axis benchmark suite has not yet emerged.
- Open scientific reproducibility is bottlenecked by the closed JFT-3B and WebLI corpora; LAION-style open data efforts must close this gap.

14.3. Future directions emerging in 2025–2026

The future directions that have emerged most recently and that we expect to dominate near-term research are:

- Hybrid Mamba-Transformer architectures that interleave selective state-space blocks with attention blocks for medium-resolution dense prediction.
- Content-adaptive tokenization that replaces fixed 16×16 patches with learned variable-resolution sampling, generalizing NaViT and the perceiver template.
- Joint generative-discriminative-contrastive pretraining that unifies MAE, DINOv2, and CLIP losses in a single objective, generalizing EVA-02 and CoCa.
- Sparse mixture-of-experts ViTs that scale total parameters to the 100B regime while keeping per-token compute bounded, building on V-MoE, LIMoE, and Soft MoE.
- Modality-universal foundation models that share a single ViT backbone across image, video, point cloud, audio, and Earth-observation modalities, generalizing DINOv2 and SAM 2.

In summary, the Vision Transformer field has matured into a structured ecosystem in which architectural family, pretraining paradigm, and downstream protocol can be selected almost orthogonally. Across these methods, the unifying empirical pattern is that combining paradigms — generative plus contrastive, plain plus adapter, attention plus state-space — beats any single family at matched compute. Crucially, the next decisive lever is data, not architecture.

15. Conclusion

This survey has reviewed Vision Transformers across thirteen substantive sections, spanning conceptual foundations, historical trajectory, architectural taxonomy, attention variants, pretraining recipes, vision-language foundation models, dense prediction, spatio-temporal and 3D extensions, datasets and benchmarks, scaling laws, robustness, application domains, and forward-looking open problems. The cumulative picture is that the Vision Transformer is no longer a competitor to convolutional networks but the default architecture of computer vision. ViT, DeiT, Swin, MAE, DINOv2, CLIP, Mask2Former, ViT-22B, SAM, and SAM 2 collectively define the canon, and ViT-5, Vision Mamba, and the next wave of multimodal foundation models extend it.

Three central tensions structure the field. First, plain versus hierarchical topology: plain ViTs win at scale, hierarchical ViTs win at moderate scale, and hybrids win at the mobile edge. Second, generative versus discriminative versus contrastive pretraining: each excels on different downstream tasks, and the strongest models combine all three. Third, softmax attention versus selective state-space alternatives: Mamba-style models look poised to dominate long-sequence vision, but softmax attention remains the moderate-sequence default.

We close with five future directions that the field will need to resolve in 2026–2027. First, content-adaptive tokenization will likely replace fixed 16×16 patches as the default. Second, hybrid Mamba-Transformer architectures will absorb the long-sequence regime. Third, joint generative-discriminative-contrastive pretraining will subsume MAE, DINOv2, and CLIP into a unified objective. Fourth, sparse mixture-of-experts will enable scaling beyond 100B parameters at bounded inference cost. Fifth, an open public corpus matching JFT-3B in scale and curation will close the reproducibility gap between academic and industrial frontier models. The architecture is now stable; the next decade of progress will be defined by data, objectives, and evaluation protocols built on top of it.

16. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. NeurIPS, 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Trans-

formers for Image Recognition at Scale. ICLR, 2021. arXiv:2010.11929.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV, 2021. doi:10.1109/ICCV48922.2021.00986.

[4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training Data-Efficient Image Transformers and Distillation Through Attention. ICML, 2021. arXiv:2012.12877.

[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. CVPR, 2022. arXiv:2111.06377.

[6] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging Properties in Self-Supervised Vision Transformers. ICCV, 2021. arXiv:2104.14294.

[7] M. Oquab, T. Darcet, T. Moutakanni, et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193, 2023.

[8] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid Vision Transformer. ICCV, 2021.

[9] W. Wang, E. Xie, X. Li, et al. PVT v2: Improved Baselines with Pyramid Vision Transformer. Computational Visual Media, 2022.

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. NeurIPS, 2021. arXiv:2105.15203.

[11] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. CVPR, 2022. arXiv:2112.01527.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-End Object Detection with Transformers. ECCV, 2020.

[13] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. ICLR, 2021.

[14] S. Zheng, J. Lu, H. Zhao, X. Zhu, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. CVPR, 2021.

[15] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid. ViViT: A Video Vision Transformer. ICCV, 2021. arXiv:2103.15691.

[16] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale Vision Transformers. ICCV, 2021. arXiv:2104.11227.

[17] Z. Tong, Y. Song, J. Wang, and L. Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. NeurIPS, 2022. arXiv:2203.12602.

[18] A. Radford, J. W. Kim, C. Hallacy, et al. Learning Transferable Visual Models From Natural Language Supervision (CLIP). ICML, 2021.

[19] J. Yu, Z. Wang, V. K. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv:2205.01917, 2022.

[20] A. Kirillov, E. Mintun, N. Ravi, et al. Segment Anything. ICCV, 2023.

[21] N. Ravi, V. Gabeur, Y.-T. Hu, et al. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714, 2024.

[22] M. Dehghani, J. Djolonga, B. Mustafa, et al. Scaling Vision Transformers to 22 Billion Parameters. ICML, 2023. arXiv:2302.05442.

[23] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. TMLR, 2022. arXiv:2106.10270.

[24] L. Yuan, Y. Chen, T. Wang, et al. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. ICCV, 2021. arXiv:2101.11986.

[25] M.-H. Guo, T.-X. Xu, J.-J. Liu, et al. Attention Mechanisms in Computer Vision: A Survey. Computational Visual Media, 2022.

[26] P. Xu, X. Zhu, and D. A. Clifton. Multimodal Learning With Transformers: A Survey. IEEE TPAMI, 2023.

[27] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapes. Video Transformers: A Survey. IEEE TPAMI, 2023.

[28] J. Mauricio, I. Domingues, and J. Bernardino. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Applied Sciences, 2023.

[29] S. Jamil, M. J. Piran, and O.-J. Kwon. A Comprehensive Survey of Transformers for Computer Vision. Drones, 2023.

[30] A. A. Aleissaee, A. Kumar, R. M. Anwer, et al. Transformers in Remote Sensing: A Survey. Remote Sensing, 2023.

[31] K. He, C. Gan, Z. Li, et al. Transformers in Medical Image Analysis. Intelligent Medicine, 2022.

[32] S. Paul and P.-Y. Chen. Vision Transformers Are Robust Learners. AAAI, 2022.

[33] Y. Bai, J. Mei, A. Yuille, and C. Xie. Are Transformers More Robust Than CNNs? arXiv:2111.05464, 2021.

[34] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh. On the Adversarial Robustness of Vision Transformers. arXiv:2103.15670, 2021.

[35] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi. Neighborhood Attention Transformer. CVPR, 2023.

[36] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu. Mobile-Former: Bridging MobileNet and Transformer. CVPR, 2022.

[37] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu. Visual Attention Network. Computational Visual Media, 2023.

[38] H. Touvron, P. Bojanowski, M. Caron, et al. ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training. IEEE TPAMI, 2022.

[39] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. arXiv:2401.09417, 2024.

[40] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s (ConvNeXt). CVPR, 2022.

[41] W. Yu, C. Si, P. Zhou, et al. MetaFormer Baselines for Vision. arXiv:2210.13452, 2022.

[42] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. arXiv:2301.00808, 2023.

[43] E. Vorontsov, A. Bozkurt, A. Casson, et al. A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection. Nature Medicine, 2024.

[44] Y. Zhou, M. A. Chia, S. K. Wagner, et al. A Foundation Model for Generalizable Disease Detection from Retinal Images (RETFound). Nature, 2023.

[45] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu. Multi-Class Token Transformer for Weakly Supervised Semantic Segmentation. CVPR, 2022.

[46] Y. Hu, Y. Cheng, A. Lu, et al. LF-ViT: Reducing Spatial Redundancy in Vision Transformer for Efficient Image Recognition. arXiv:2402.00033, 2024.

[47] Y. Xu, Z. Zhang, M. Zhang, et al. Evo-ViT: Slow-Fast Token Evolution for Dynamic Vision Transformer. AAAI, 2022. arXiv:2108.01390.

[48] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang. Vision Transformers for Image Classification: A Comparative Survey. Technologies, 2025.

[49] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei. MOTR: End-to-End Multiple-Object Tracking with Transformer. ECCV, 2022.

[50] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. DETRs Beat YOLOs on Real-time Object Detection (RT-DETR). CVPR, 2024.

[51] X. Li, H. Ding, H. Yuan, et al. Transformer-Based Visual Segmentation: A Survey. arXiv:2304.09854, 2023.

[52] H. Cao, Y. Wang, J. Chen, et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. ECCV Workshops, 2023.

[53] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. MICCAI, 2022.

[54] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Al Dayil, and N. Al Ajlan. Vision Transformers for Remote Sensing Image Classification. Remote Sensing, 2021.

[55] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon. A Survey on Masked Autoencoder for Self-Supervised Learning in Vision and Beyond. arXiv:2208.00173, 2022.

[56] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment Anything in Medical Images (MedSAM). Nature Communications, 2024.

[57] F. Wang, S. Ren, T. Zhang, et al. ViT-5: Vision Transformers for The Mid-2020s. arXiv:2602.08071, 2026.

[58] W. Shi, J. Xu, and P. Gao. SSformer: A Lightweight Transformer for Semantic Segmentation. MMSP, 2022.

[59] K. Al-Hammuri, F. Gebali, A. Kanan, and I. T. Chelvan. Vision Transformer Architecture and Applications in Digital Health: A Tutorial and Survey. Visual Computing for Industry, Biomedicine, and Art, 2023.

[60] W. Wu, Z. Sun, and W. Ouyang. Revisiting Clas-

sifier: Transferring Vision-Language Models for Video Recognition. AAAI, 2023.

[61] S. Lavoie, P. Kirichenko, M. Ibrahim, M. Assran, A. G. Wilson, A. Courville, and N. Ballas. Modeling Caption Diversity in Contrastive Vision-Language Pretraining. arXiv:2405.00740, 2024.

[62] X. Yang and X. Wang. Kolmogorov-Arnold Transformer. arXiv:2409.10594, 2024.

[63] L. Melas-Kyriazi. Do You Even Need Attention? A Stack of Feed-Forward Layers Does Surprisingly Well on ImageNet. arXiv:2105.02723, 2021.

[64] A. Hassani and H. Shi. Dilated Neighborhood Attention Transformer. arXiv:2209.15001, 2022.

[65] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. arXiv:2209.08575, 2022.

[66] K. Han, J. Guo, Y. Tang, and Y. Wang. PyramidTNT: Improved Transformer-in-Transformer Baselines with Pyramid Architecture. arXiv:2201.00978, 2022.

[67] W. Sun, Z. Qin, H. Deng, et al. Vicinity Vision Transformer. arXiv:2206.10552, 2022.

[68] Q. Zhang, Y. Xu, J. Zhang, and D. Tao. ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond. International Journal of Computer Vision, 2023. arXiv:2202.10108.

[69] B. Cheng, A. G. Schwing, and A. Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation (MaskFormer). NeurIPS, 2021.

[70] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li. Large Selective Kernel Network for Remote Sensing Object Detection. arXiv:2303.09030, 2023.

[71] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. ECCV, 2022.

[72] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. PCT: Point Cloud Transformer. Computational Visual Media, 2021.

[73] Y. Pang, W. Wang, F. E. H. Tay, W. Liu, Y. Tian, and L. Yuan. Masked Autoencoders for Point Cloud Self-Supervised Learning (Point-MAE). arXiv:2203.06604, 2022.

[74] X. Chen, J. Djolonga, P. Padlewski, et al. PaLI-X: On Scaling up a Multilingual Vision and Language Model. arXiv:2305.18565, 2023.

[75] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. CVPR, 2009.

[76] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. ECCV, 2014.

[77] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene Parsing through ADE20K Dataset. CVPR, 2017.

[78] M. Cordts, M. Omran, S. Ramos, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. CVPR, 2016.

[79] W. Kay, J. Carreira, K. Simonyan, et al. The Kinetics Human Action Video Dataset. arXiv:1705.06950, 2017.

[80] C. Schuhmann, R. Beaumont, R. Vencu, et al. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. NeurIPS Datasets Track, 2022.

[81] D. Hendrycks, S. Basart, N. Mu, et al. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. ICCV, 2021.

[82] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet Classifiers Generalize to ImageNet? ICML, 2019.

Term	Definition
Attention rollout	Iterative composition of attention matrices across layers to attribute output to input tokens.
BEV	Bird's-eye view, a top-down 2D feature map used in autonomous driving.
Bipartite matching	Hungarian-algorithm assignment used by DETR to match predicted to ground-truth objects.
[CLS] token	Learnable classification token prepended to a ViT input sequence.
Cross-attention	Attention where queries come from one sequence and keys/values from another.
Deformable attention	Sparse attention with learned reference-point offsets (Deformable DETR).
Distillation token	DeiT's auxiliary token trained to match teacher predictions via hard distillation.
Hungarian matching	Polynomial-time algorithm for optimal one-to-one assignment.
LayerNorm	Per-token feature normalization used in Transformer blocks.
LoRA	Low-Rank Adaptation, a parameter-efficient fine-tuning method.
Masked image modeling (MIM)	Self-supervised pretraining by reconstructing masked patches (MAE, SimMIM, BEiT).
Multi-head self-attention (MSA)	Parallel attention heads each computing $\mathrm{softmax}(QK^\top/\sqrt{d_k})V$.
NaViT	Native-aspect-ratio Transformer that packs multiple variable-resolution images.
NMS	Non-Maximum Suppression, the post-hoc filter eliminated by DETR.
Patch tokenization	Splitting an image into a grid of patches, each linearly embedded as a token.
Patch embedding	The learned linear projection $E \in \mathbb{R}^{P^2C \times D}$ mapping patches to tokens.
PQ	Panoptic Quality, the joint metric for semantic + instance segmentation.
Pre-norm	Layer-normalization placement before attention/MLP rather than after.
Prompt encoder	SAM's module that encodes points/boxes/text prompts as embeddings.
Q-Former	BLIP-2's bridging Transformer between frozen ViT and frozen LLM.
Register tokens	Auxiliary tokens that absorb attention-sink artifacts (Darcet 2024).
Relative position bias	Position-aware bias added inside attention (Swin, NAT).
Self-distillation	Student-teacher SSL where teacher is an EMA of student (DINO).
SA-1B	Segment Anything 1-Billion: SAM's training corpus (11M images, 1.1B masks).
Set prediction	DETR's objective: predict an unordered set of objects via Hungarian loss.
Shifted window	Swin's mechanism for cross-window attention via cyclic shift.
Spatial-reduction attention (SRA)	PVT's downsampled-K/V attention for global mixing at low resolution.
Stochastic depth	Per-layer dropout that skips entire residual blocks, used in DeiT/Swin.
Token merging (ToMe)	Bipartite soft matching of similar tokens to reduce sequence length.
Token pruning	Discarding low-importance tokens (DynamicViT, Evo-ViT).
Tube masking	VideoMAE's strategy of masking spatially aligned cubes across frames.
ViT-22B	The 21.7B-parameter ViT (Dehghani et al., 2023).
Window attention	Self-attention restricted to local non-overlapping windows (Swin).
Zero-shot	Inference on a task without any task-specific training (CLIP-style).