WHAT HAVE VISION MODELS LEARNED IN THE DEEP LEARNING ERA?

Anonymous authors
Paper under double-blind review

ABSTRACT

What have vision models learned in the deep learning era, and how general are their representations? We systematically revisit the decade of visual self-supervised learning (VSSL) and geometry, generation, and video models, with references including supervised models, vision-language models (VLMs), multimodal world models (MMWMs), and frontier models. Through over 2,500 training and evaluation experiments, we distinguish original-setting comparisons, controlled pre-training, and common downstream evaluation on the Basic Five: image classification, object detection, segmentation, depth estimation, and video classification, complemented by an extended transfer suite. All scores are computed from our own environment, never copied from prior reports. Our main findings connect the past, present, and future directions of visual representation learning. (Past) Over the past decade, vision-native representations have become substantially more general, supporting transfer across semantics, dense prediction, and geometry, although progress and model strengths remain task-dependent. (Present) Modern vision-native representations, exemplified by DINOv3, outperform the evaluated VLM and MMWM references on multiple visual tasks, demonstrating broad transfer from vision-native learning itself. (Future direction) These achievements are not fixed endpoints: learning-design changes improve both classical and modern VSSL approaches, while scaling yields further gains across all five core tasks. Moreover, a separate matched-subset comparison with GPT-6 Astra reveals complementary strengths: DINOv3 performs better on ADE20K semantic-point accuracy and comparably on NYUv2 relative-depth accuracy. Overall, these results connect historical progress to the current frontier and future directions, providing an empirical account of what vision models have learned and a foundation for developing more general vision-native representations.[1]

1 INTRODUCTION

Since ILSVRC 2012 (Russakovsky et al., 2015), vision models have expanded from image classification to detection, segmentation, 3D structure estimation, generation, and video understanding. Yet acquiring new functions does not necessarily produce representations reusable beyond those functions. As visual research diversifies, its advances need to be examined through this shared perspective. We mainly focus on vision-native representation learning across VSSL, geometry, generation, and video models. Supervised models, VLMs, and MMWMs serve as references. Here, vision-native learning constructs pre-training signals from images, videos, or geometric observations without directly using manually defined semantic category labels or language-derived inputs or supervision. We connect the past, present, and future of vision-native representation learning through a systematic investigation of the deep learning era.

Past in vision-native models (Section 3). Figure 1 traces visual transfer across model families and tasks using downstream scores measured in our environment, rather than compiled from prior reports. This retrospective addresses our central question: what have vision-native models learned in the deep learning era, and how broadly do their representations support other applications? The

[1] Anonymized code for reproducing the training and evaluation procedures is available at `https://anonymous.4open.science/r/review-artifact-2027-B703/`. Scores are reported in the paper and supplementary materials; our trained checkpoints and complete experiment logs will be released.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

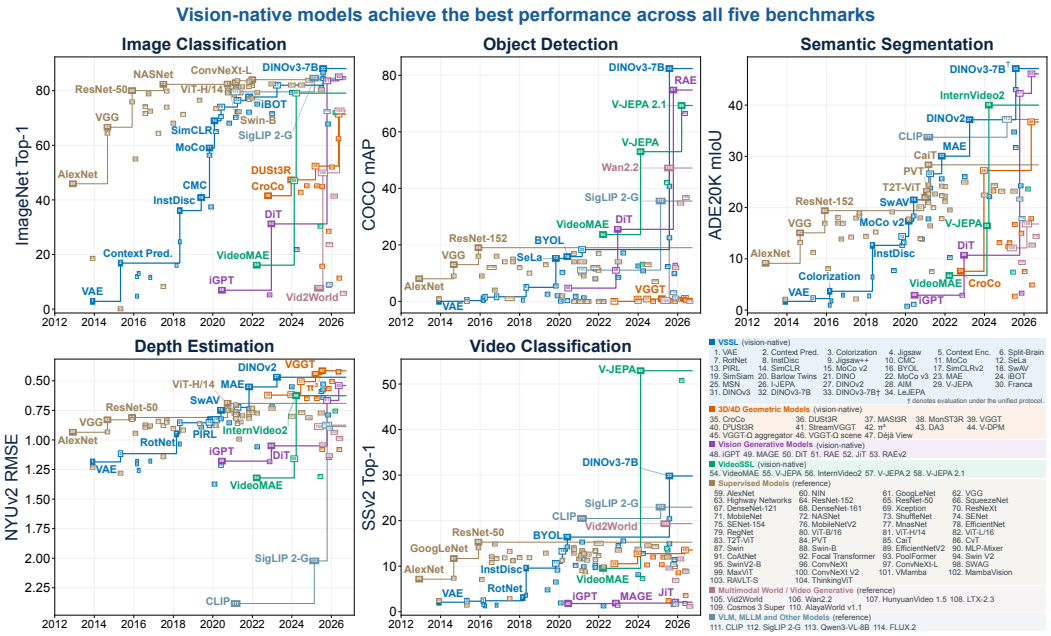


Figure 1: Historical transfer across the Basic Five, measured in our environment rather than copied from prior reports. Models are ordered by first public release; running-best curves summarize each family within its evaluation series. Models are evaluated using category-specific feature extraction and downstream protocols. Complete and more readable benchmark-wise plots appear in Appendix C.

Basic Five including image classification, object detection, semantic segmentation, depth estimation, and video classification, anchors these comparisons, complemented by broader transfer tasks.

Present in vision-native models (Section 4). Figure 2 summarizes transfer across an extended suite beyond Basic Five. Among the evaluated models, DINOv3-7B (Siméoni et al., 2026) ranks first on 31 benchmarks and second on 6, demonstrating broad transfer. These results show that vision-native representations can support diverse applications without language supervision, while leaving distinct strengths in other models. Generality therefore requires examining the breadth of transfer alongside task-specific advantages, rather than relying on a single benchmark.

Future directions for vision-native models (Section 5). Revisiting existing learning designs and model scales reveals further opportunities for improvement. A separate comparison with GPT-6 Astra (OpenAI, 2026b) and GPT-5.6 Sol (OpenAI, 2026a), using matched samples and GPT-adapted task interfaces, provides a frontier reference: DINOv3-7B exceeds both systems on ADE20K and performs comparably on NYUv2. The strong transfer of representations learned from still images alone across diverse visual tasks demonstrates the broad utility of vision-native models and suggests opportunities to extend their capabilities by also learning from geometric and spatiotemporal information.

Connecting past achievements, present capabilities, and future directions, this study motivates integrating recognition, geometry, generation, and video to develop representations whose usefulness grows across tasks. Our contribution is not a new learning algorithm, but an empirical foundation through reproducible comparisons and diagnostic experiments. These findings point toward generalist vision-native models that integrate new capabilities while preserving reusable features. Model configurations and complete results appear in Appendices B and C, respectively.

2 Models and Evaluation Framework

Vision-native and reference models. We define strict vision-native models as those pre-trained simply from visual modalities, including images, videos, and geometric information, without using

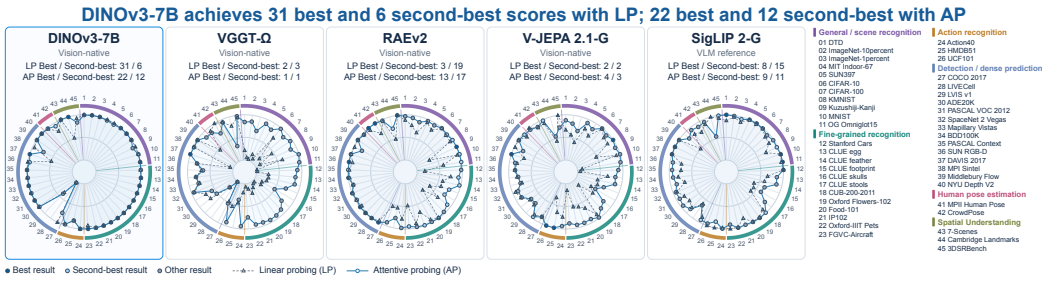


Figure 2: Broad transfer of vision-native models. Radial values are scaled to $[0, 1]$ for visualization, with larger values indicating better performance. Non-error scores are divided by 100, while errors e are transformed as $1/(1 + e)$. For each benchmark and probing method (LP/AP), values are then divided by the maximum across the five models. DINOv3-7B ranks first on 31 of 45 benchmarks and second on 6 with LP. SigLIP 2-G serves as a VLM reference.

image-category labels, text or language-derived features, audio, or action as direct inputs or learning targets. This category applies to evaluated checkpoints and components (Appendix B).

Following Figure 1, we group primary models into VSSL, 3D/4D Geometric Models, Vision Generative Models, and Video SSL. VSSL provides the main setting for tracing changes in visual learning objectives. The other groups extend this investigation to 3D/4D geometry, visual generation, and learning from video. Evaluated checkpoints or components qualify as strict vision-native when their pre-training inputs and targets satisfy the above definition. These groups describe different learning objectives and visual capabilities, not mutually exclusive categories or successive replacements.

ImageNet-supervised models, VLMs, and MMWMs serve as references for assessing vision-only learning against other forms of supervision. Supervised models connect this comparison to image recognition and transfer learning (Deng et al., 2009; Russakovsky et al., 2015; Huh et al., 2016; Kornblith et al., 2019), while VLMs and MMWMs provide language-aligned and multimodal references. We evaluate specified visual components, not full-system capabilities. GPT models provide a separate system-level reference under adapted task interfaces.

Core benchmarks and feature evaluation. The Basic Five (Figure 1) comprises ImageNet-1K (Russakovsky et al., 2015), COCO (Lin et al., 2014b), ADE20K (Zhou et al., 2017), NYUv2 (Silberman et al., 2012), and Something-Something V2 (SSv2) (Goyal et al., 2017). We select these tasks to test complementary uses of image-level recognition, object localization, pixel-level labeling, geometric estimation, and temporal recognition. The Basic Five provides a shared set of tasks and metrics for comparing vision-native models as consistently as possible beyond classification. These tasks offer an initial assessment, not an exhaustive measure of visual capabilities.

Our default evaluation freezes the pre-trained feature extractor and trains only a task-specific predictor, testing how readily its features support downstream tasks. We favor simple predictors applicable across model families, using a linear classifier for image classification and task-appropriate heads for structured outputs such as detection. We therefore describe this setting as frozen-feature probing, rather than treating every task as linear probing. Within each comparison, we specify feature extraction and align task-specific evaluation rules. We additionally use attentive probing (AP) to examine a more expressive use of frozen features, and fine-tuning (FT) to assess adaptation when the backbone is also updated. AP results appear in the main paper and FT results in the Appendix C.

Representative selection and broader evaluation. We select representative models using the Basic Five. Within each category, selection considers performance across tasks and task-specific strengths under specified feature extraction, retaining multiple candidates when results are close or complementary. We report each configuration rather than combine task-wise best scores into a single model.

3 Past: Progress and Transfer in Vision-Native Models

Figure 1 provides a historical overview of visual models. This section examines their transfer performance under original training settings and category-specific evaluation.

Original-setting evaluation (ASIS). Under ASIS, we evaluate released checkpoints or reproduce pre-training following the original architecture and training procedure. Figure 1 summarizes their transfer across the Basic Five under category-specific evaluation; pre-training scale and downstream conditions are not fully aligned across categories. Section 4 provides unified downstream comparisons, while Section 5.1 returns to the VSSL ASIS results as a reference for controlled pre-training. Checkpoint provenance, deviations from original settings, and complete results are documented in Appendices B and C. When reporting five-score tuples, we use the order {ImageNet-1K Top-1, COCO detection mAP, ADE20K mIoU, NYUv2 RMSE, SSv2 Top-1}.

VSSL: broader transfer with task-specific strengths. We evaluate 32 VSSL methods spanning early and contemporary learning approaches with available implementations or checkpoints. Under ASIS, ImageNet-1K accuracy ranges from 2.83% for the evaluated VAE (Kingma & Welling, 2014) to 84.47% for DINOv3 ViT-B/16 (Siméoni et al., 2026). DINOv3 achieves $\{84.47, 22.50, 43.06, 0.528, 20.41\}$, leading on four of the five tasks among the evaluated VSSL methods (Table 10). DINOv2 (Oquab et al., 2024) nevertheless retains an advantage in depth estimation, with an RMSE of 0.470. Its strengths extend beyond image classification to dense prediction and video classification, illustrating the breadth of representations learned from visual observations. These results show transfer beyond image classification, although no single method leads on every task. CTRL revisits these ASIS comparisons under aligned pre-training conditions.

3D/4D Geometric Models: strengths depend on the representation. We compare models of static and time-varying geometry, distinguishing aggregated image features, scene tokens, and spatial features. The VGGT-Ω (Wang et al., 2026a) aggregator without text alignment achieves $\{71.41, 1.06, 17.84, 0.427, 13.54\}$, whereas its scene representation achieves $\{62.39, 0.62, 36.73, 0.421, 8.96\}$. The aggregator is stronger in image and video classification, while scene features better support segmentation and depth estimation. Meanwhile, π^3 (Wang et al., 2026b) achieves $\{44.83, 0.02, 16.71, 0.415, 12.08\}$, yielding the lowest depth RMSE among the geometric configurations shown in Figure 1. Transfer strengths therefore depend on the extracted representation, not simply the model family.

Vision Generative Models: which components support transfer? We evaluate encoder, tokenizer, and generator features separately, measuring transfer rather than generation quality. The RAEv1 (Zheng et al., 2026) encoder achieves $\{83.60, 74.80, 42.30, 0.666, 1.60\}$, while RAEv2 (Singh et al., 2026) generator features from block 29 at $t = 0.1$ achieve $\{85.10, 66.50, 46.10, 0.540, 1.80\}$. These configurations provide strong image-level and pixel-level features but limited video-classification performance under the evaluated settings. Component and timestep ablations distinguish inherited encoder representations from features accessed within the generator (Table 28).

Video SSL: transfer from video to image tasks. Video SSL learns from temporal order (Misra et al., 2016), missing content (Tong et al., 2022), or latent prediction, including the V-JEPA series (Bardes et al., 2024; Assran et al., 2025). The V-JEPA 2.1 (Mur-Labadia et al., 2026) ViT-g configuration shown in Figure 1 achieves $\{71.95, 69.33, 23.63, 0.756, 50.80\}$. The ViT-L configuration achieves $\{72.64, 65.95, 23.19, 0.786, 49.85\}$ (Table 29). ViT-g performs better on four tasks, whereas ViT-L has higher image-classification accuracy. These results demonstrate reuse of video-learned features across image and video tasks, with task-dependent differences between configurations. Frame sampling and feature aggregation are documented, and action-conditioned or language-based configurations are treated separately as references.

Reference models: VLMs and Multimodal World Models. We select SigLIP 2 Giant (Tschannen et al., 2025) and Cosmos 3 Super (64B) (NVIDIA, 2026) as references for language-aligned and multimodal world-model learning, respectively. These contrasting profiles motivate their inclusion as reference points, not as uniformly superior models within their categories. We evaluate specified visual components rather than whole-system generation or reasoning capabilities, with component and feature-extraction ablations reported in Appendix C.

4 PRESENT: UNIFIED AND EXTENDED EVALUATION

We compare DINOv3-7B, the VGGT-Ω aggregator without text alignment, RAEv2-K7, and V-JEPA 2.1 ViT-g against SigLIP 2 Giant and the visual encoder and merger of Cosmos 3 Super under

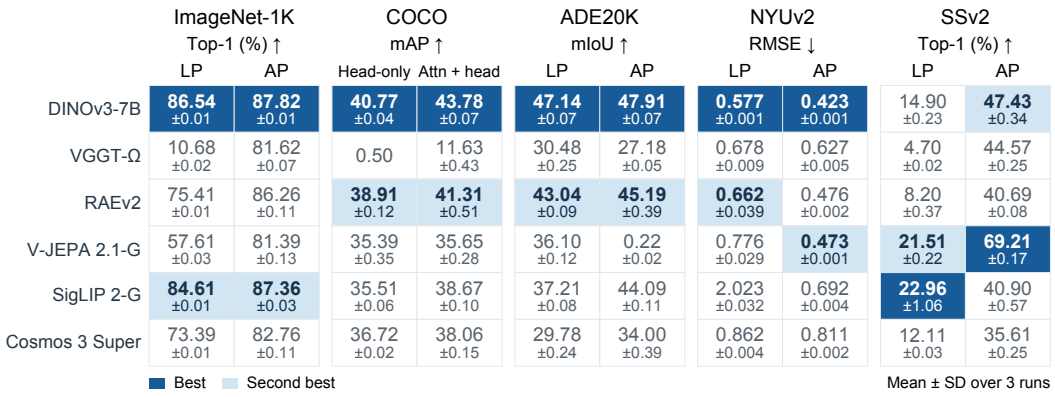


Figure 3: Broad transfer depends on both the representation and how it is used. Basic Five performance under frozen-feature probing (LP) and attentive probing (AP). Shading indicates the best and second-best scores among the displayed models within each task and protocol. Results are reported as mean ± standard deviation over three seeds where available; repeat counts appear in Appendix C.

common downstream conditions. We apply repeat unified Basic Five LP/AP evaluation with three random seeds (Appendix E), reporting the mean and standard deviation.

Unified evaluation settings. We compare frozen-feature probing and attentive probing (AP) using task-specific predictors and attention-based predictors, respectively, while keeping the backbone frozen. Downstream conditions are aligned across models under unified Basic Five LP/AP evaluation, but pre-training data and model scale are not matched. Optimization and augmentation may differ between tracks, so their comparison does not isolate predictor design alone. Complete results, repeat counts, and variability are reported in Appendix C, including FT results in Table 38.

Basic Five: broad transfer and protocol-dependent strengths. Figure 3 compares frozen-feature probing and attentive probing (AP). Under frozen-feature probing, DINOv3-7B achieves $\{86.54, 40.77, 47.14, 0.577, 14.90\}$, outperforming SigLIP 2 Giant's $\{84.61, 35.51, 37.21, 2.023, 22.96\}$ on four tasks. These results demonstrate broad transfer from still-image pre-training to classification, localization, segmentation, and depth estimation, but weaker transfer to video classification. Under AP, DINOv3-7B reaches $\{87.82, 43.78, 47.91, 0.423, 47.43\}$, improving across all five tasks and overtaking SigLIP 2 Giant on SSv2, where the latter achieves 40.90%. V-JEPA 2.1 nevertheless leads the evaluated representatives on SSv2 under AP at 69.21%, highlighting the complementary strengths of video-native learning. Thus, broad transfer and task-specific advantages coexist, and model rankings depend on how frozen representations are used. These comparisons characterize the evaluated protocols rather than each model's optimal performance. Complete results and variability appear in Tables 36 and 37, with FT results provided separately in Table 38.

Extended evaluation: broad transfer and complementary strengths. Figure 2 extends the comparison beyond Basic Five to fine-grained recognition, texture and scene recognition, video understanding, and other applications. Vision-native models provide broadly reusable features, with DINOv3-7B showing strengths across diverse recognition tasks and specialized models retaining advantages in segmentation and video classification. Relative to the evaluated references, these results show that language supervision is not required for strong transfer across diverse visual applications. However, advantages are not uniform within task categories. The VLM reference remains competitive and leads on some fine-grained and scene-recognition benchmarks. These findings concern the extracted visual features, not the capabilities of the complete systems. We therefore assess generality through both transfer breadth and complementary strengths, using shared valid entries for aggregate comparisons. Complete scores appear in Appendix C.

Lessons from both evaluations: broad transfer and protocol-dependent strengths. Across Basic Five and extended evaluation, vision-native representations compete strongly with language-supervised and multimodal references while specialized models retain complementary strengths, notably V-JEPA 2.1 under attentive probing. Protocol-dependent rankings show why a single frozen-

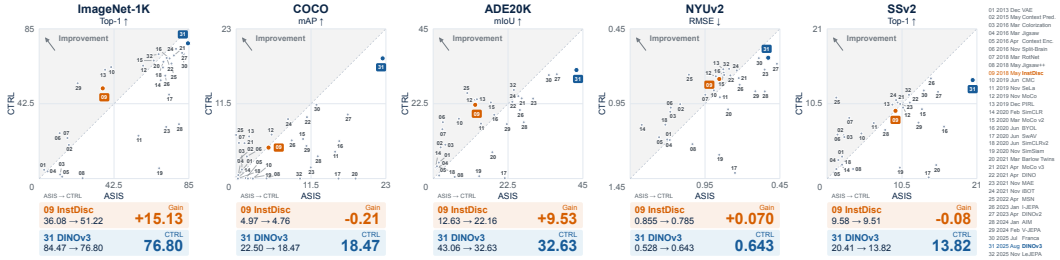


Figure 4: Performance of 32 VSSL methods on the Basic Five under ASIS and CTRL. Despite extensive study of contrastive learning and later methods, earlier approaches merit revisiting: Instance Discrimination gains 15.13 points on ImageNet-1K. DINOv3 leads on four tasks under CTRL. Both retain room for learning-design improvements (Section 5.2). Gray regions indicate gains over ASIS.

feature leaderboard cannot capture representation utility. Shared downstream conditions enable comparison but may suit models differently and do not establish performance ceilings. Cross-track differences likewise cannot be attributed solely to predictor design when optimization and augmentation differ. Transfer breadth and task-specific strengths must therefore be assessed under explicit evaluation conditions. Broad utility alone does not establish improvability.

5 Future: Exploring Further Progress in Vision-Native Learning

Can broadly transferable representations improve further? We investigate this through CTRL and changes to learning design. We also include a frontier-model comparison as a reference.

5.1 CTRL: Revisiting VSSL under Controlled Conditions

Comparison between ASIS and CTRL (Figure 4). To compare VSSL in CTRL conditions, we revisit the 32 VSSL methods evaluated under ASIS in Section 3. Each method is pre-trained with ViT-B/16 on unlabeled ImageNet-1K for 300 epochs, with checkpoints from the same training run evaluated at epochs 100, 200, and 300. We use common optimization and augmentation settings where compatible with each method, while preserving defining mechanisms such as masking, cropping, and teacher updates. CTRL therefore compares how existing methods perform under the shared pre-training framework, rather than isolating the learning objective alone.

Revisiting both early and contemporary approaches. Figure 4 compares ASIS with CTRL at 300 epochs. Instance Discrimination improves ImageNet accuracy by 15.13 percentage points over its ASIS reference. Earlier reported results therefore need not define the limits of these learning approaches. These gains reflect their behavior under changed training and evaluation conditions, rather than improvements attributable to a new learning objective. Section 5.2 examines whether additional changes can extend these gains.

DINOv3 achieves the highest scores on classification, detection, segmentation, and depth estimation among the evaluated VSSL methods under CTRL, with Basic Five scores of $\{76.80, 18.47, 32.63, 0.643, 13.82\}$ (Table 13). These results are obtained with ViT-B/16 pre-trained on unlabeled ImageNet-1K, without the billion-scale pre-training data used for its released checkpoints. This comparison assesses compatibility with the shared training framework, not the intrinsic superiority of an objective or the best attainable performance of every method. The early and contemporary examples motivate the separate diagnostic experiments in Section 5.2.

5.2 An Empirical Exploration of Future Directions

From the comparison between ASIS and CTRL, we examine learning design changes in Instance Discrimination and DINOv3 as diagnostic experiments, not novel methods. Both use ViT-B/16 pre-trained on ImageNet-1K for 300 epochs, with an additional ViT-L/16 multi-crop run for Instance Discrimination. At the same time, additional experiments for Rotation, Colorization, and SplitBrain appear in Appendix C (Tables 21, 20, and 22).

Instance Discrimination: revisiting views and model capacity. Figure 5 compares single-view ViT-B/16, multi-crop ViT-B/16, and multi-crop ViT-L/16, all pre-trained on ImageNet-1K for 300 epochs. Multi-crop training retains the original objective and memory bank while using two global and four local views per image. From single-view ViT-B to multi-crop ViT-L, Basic Five scores change by $\{+21.28, +3.09, +10.59, -0.073, +2.89\}$, improving all five tasks. These gains illustrate opportunities from revisiting views and capacity within an early learning approach, but the additional computation precludes claims of improved compute efficiency.

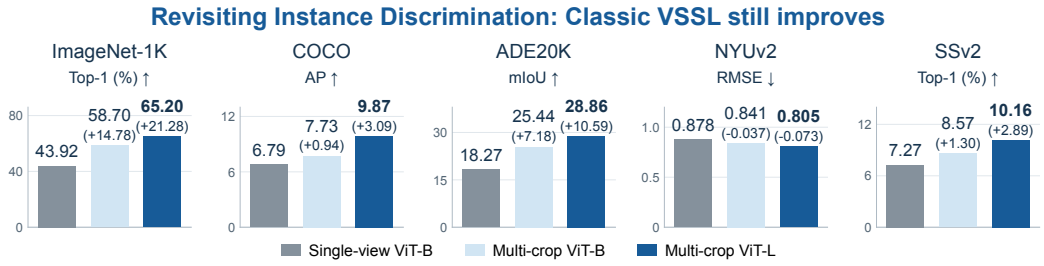


Figure 5: Revisiting Instance Discrimination improves all five tasks.

DINOv3: a diagnostic study of shared head. We examine a shared head for image-level DINO and patch-level iBOT learning together with a broader masking range, as a diagnostic experiment rather than a novel method. Relative to the original separate-head configuration, Basic Five scores change by $\{+0.59, +1.16, +0.32, +0.001, -0.20\}$, improving classification, detection, and segmentation while slightly worsening depth RMSE and video Top-1 accuracy. Secondary metrics show gains of 0.40 points in ImageNet Top-5 and 0.51 points in SSv2 Top-5, alongside a 0.002 improvement in NYUv2 AbsRel. These single-run results suggest opportunities to revisit contemporary learning designs, while revealing task-dependent trade-offs rather than uniform improvement.

Model scaling: continued gains at the largest evaluated scale. We compare DINOv3 checkpoints at ViT-B, L, H, and 7B scales. The 7B model performs best on all five tasks among these checkpoints. From ViT-H to 7B, Basic Five scores change from $\{86.31, 36.43, 45.55, 0.700, 14.22\}$ to $\{86.54, 40.77, 47.14, 0.577, 14.90\}$, corresponding to $\{+0.23, +4.34, +1.59, -0.123, +0.68\}$. Thus, modest classification gains accompany further improvements in detection, segmentation, depth estimation, and video classification. Although intermediate scales are not uniformly ordered, gains continue between the two largest evaluated checkpoints rather than exhibiting a common plateau across tasks. These observations motivate exploring scales beyond 7B, without establishing a scaling law or guaranteeing further gains. Complete results appear in Tables 17 and 36.

Frontier-model reference (Figure 6). For reference, we compare DINOv3-7B with frontier models using GPT-6 Astra and GPT-5.6 Sol, separately from the original Basic Five metrics (details in Appendix C.15). Against Astra, DINOv3 achieves higher ADE20K semantic-point accuracy (79.99% LP / 80.98% AP versus 66.96%) and similar NYUv2 relative-depth accuracy (80.30% LP versus 80.02%), while Astra leads on ImageNet-1K, COCO, and SSv2. This frontier model comparison highlights complementary strengths, not equivalent training conditions. DINOv3's still-image pretraining already supports these capabilities; jointly learning geometry and spatiotemporal structure remains a promising direction for extending vision-native representations.

6 Visual Models through Representations and Transfer

A representation-based map of visual models (Figure 7(a)). We map the evaluated models according to the inter-image similarity structures induced by their representations. Adapting the *Map of Encoders* framework (Zhang & Bollegala, 2026), we construct quantum relative entropy (QRE) descriptors from features extracted on all 50,000 ImageNet-1K validation images and project them into two dimensions using t-SNE. The resulting map provides a qualitative overview of representational relationships without using downstream performance scores.

A performance-based map of visual models (Figure 7(b)). We apply PCA to the Basic Five score vectors, with NYUv2 RMSE sign-reversed. The horizontal axis (PC1; 47.49% of variance) com-

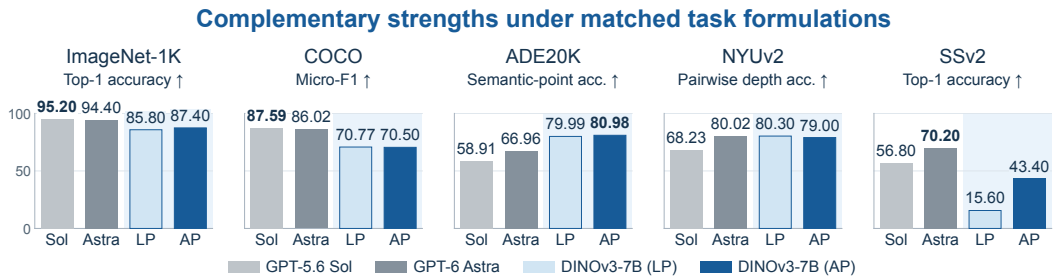


Figure 6: Reference-only comparison of DINOv3-7B, GPT-6 Astra, and GPT-5.6 Sol using GPT-adapted task formulations on matched samples from the matched held-out subsets. DINOv3-7B (LP) and Astra show comparable NYUv2 accuracy. Training conditions and evaluation interfaces differ.

bines all five metrics with positive weights, summarizing overall Basic Five performance, while the vertical axis (PC2; 21.48%) contrasts relative strengths in object detection and video classification with those in image classification and depth estimation under the evaluated settings. These complementary transfer profiles suggest a potential next frontier for vision-native models: integrating recognition, geometry, and spatiotemporal learning within a broadly reusable representation. A key research goal is to improve transfer across tasks while preserving existing strengths, rather than merely shifting task preferences. The map also reveals models that stand out in overall performance.

7 Related Work

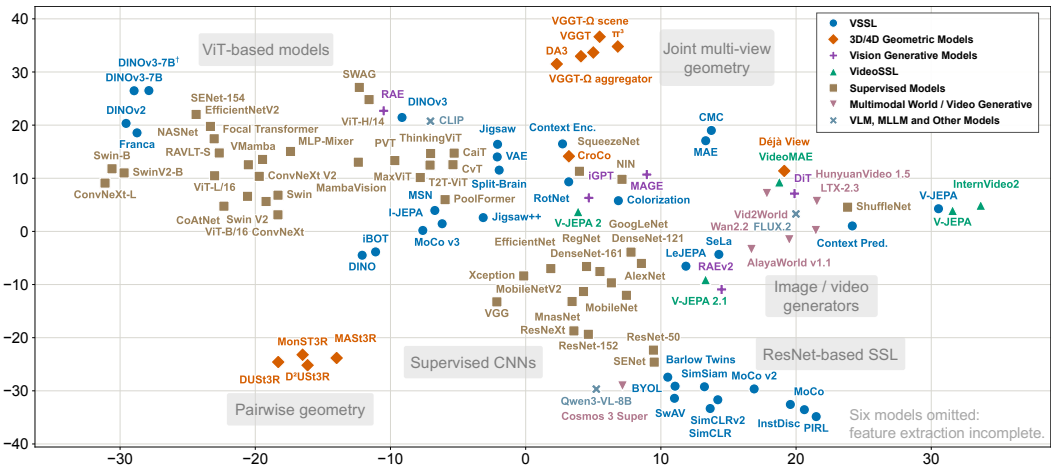
Transfer breadth: toward generalist visual representations. ImageNet studies on data configuration, performance, and evaluation help distinguish recognition gains from progress in transferable representations (Huh et al., 2016; Kornblith et al., 2019; Beyer et al., 2020). Evaluations across diverse tasks and domains similarly assess utility beyond a single benchmark (Zhai et al., 2019; Ericsson et al., 2021; Islam et al., 2021). We extend this perspective to geometric, generative, and video models, with VLMs and MMWMs as references. Basic Five and extended evaluation examine both transfer breadth and dependence on how features are used. Our focus is the reuse of visual representations, rather than the complete systems' original capabilities.

Controlled pre-training: revisiting methods under shared conditions. VSSL reevaluations highlight the importance of architecture and pre-training data scale in method comparisons (Kolesnikov et al., 2019; Goyal et al., 2019). We pair original-setting evaluation (ASIS) with shared pre-training conditions (CTRL) across early and contemporary methods. This distinguishes historical achievements from performance within a common framework, without attributing differences solely to learning objectives or treating shared settings as optimal for every method.

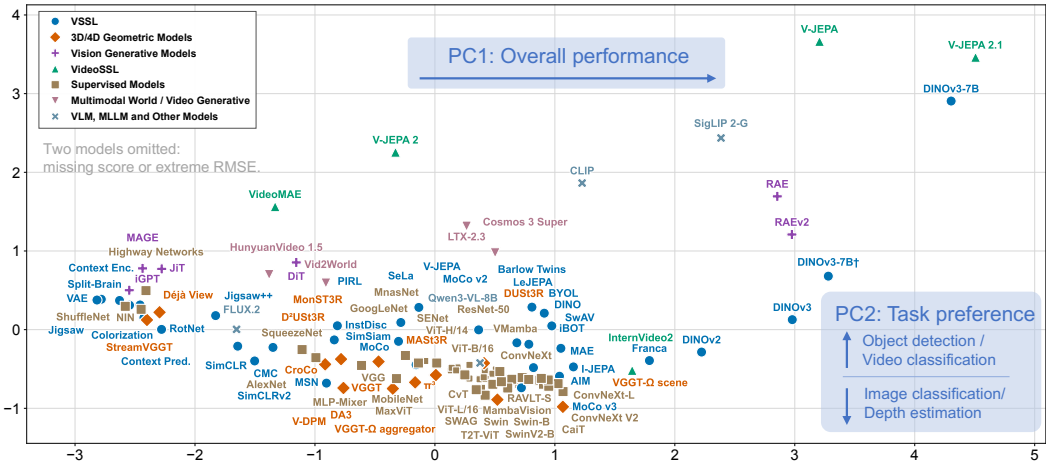
Future directions: revisiting learning principles and model scale. Instance Discrimination, rotation prediction, and DINOv3 exemplify different ways to construct learning targets from visual observations (Wu et al., 2018; Gidaris et al., 2018; Siméoni et al., 2026). Our diagnostic experiments revisit existing designs and compare released model scales, examining further gains and task-dependent trade-offs rather than proposing a new learning algorithm. A separate frontier-system comparison contextualizes the capabilities accessible from vision-native representations. These investigations motivate studying how geometric and spatiotemporal learning can extend broadly transferable features while preserving their existing utility.

8 Discussion and Conclusion

We revisit vision models across the deep learning era, compare contemporary representations across diverse transfer tasks, and explore future opportunities through diagnostic experiments on existing approaches. The results reveal broad transfer from vision-native learning, complementary strengths across model families, and further gains within the evaluated learning designs and model scales.



(a) Representation similarity map.



(b) Basic Five performance map.

Figure 7: Representation- and performance-based maps of visual models. (a) t-SNE of QRE descriptors computed from ImageNet-1K validation features. (b) PCA of Basic Five scores, with NYUv2 RMSE sign-reversed so that higher values consistently indicate better performance. Each point represents an evaluated model configuration; colors and marker shapes indicate model categories.

Progress is not fully captured by classification accuracy or the acquisition of new functions. Vision-native representations support broad transfer, but complementary strengths and protocol-dependent rankings argue against a single universal leaderboard. Gains across all five tasks from DINOv3 ViT-H to 7B, together with gains and trade-offs from revisiting early and contemporary learning designs, identify concrete opportunities for further development. The frontier-system comparison also highlights capabilities already accessible from still-image representations.

These conclusions concern the evaluated tasks and protocols, not whole-system capabilities or performance ceilings. Shared conditions may suit models differently, while differences in pre-training data, scale, and system interfaces limit causal attribution. Task-specific fine-tuning does not establish simultaneous capability retention. Broader coverage and repeated runs remain necessary, particularly for single-run design experiments.

A promising direction is to incorporate geometry, generation, and spatiotemporal learning while preserving broadly reusable features. Evaluating matched checkpoints before and after learning on both new functions and existing transfer tasks would directly test acquisition and retention. Generalist vision-native models remain an open objective, but broad transfer and experimentally demonstrated opportunities for further gains make their development an increasingly tangible research direction.

REPRODUCIBILITY STATEMENT

Anonymized code for reproducing our training and evaluation procedures is available at the link: https://anonymous.4open.science/r/review-artifact-2027-B703/. Model configurations and complete scores are documented in the appendices. We plan to release our trained checkpoints and complete experiment logs to support further reproduction and independent verification.

AI USE STATEMENT

Generative AI tools were used to support manuscript preparation and polishing, minor error correction, code generation, and research execution workflows. The authors reviewed the resulting text and suggestions and take responsibility for the experiments, analyses, and final manuscript.

REFERENCES

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Interactive long-horizon world modeling – full technical report (v1.1). *arXiv preprint arXiv:2608.13492*, 2026. URL https://arxiv.org/abs/2608.13492.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/papers/Andriluka_2D_Human_Pose_2014_CVPR_paper.pdf.

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyx-jyBFPr.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, 2022. URL https://link.springer.com/chapter/10.1007/978-3-031-19821-2_26.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/baevski23a.html.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision*, 2011. URL https://doi.org/10.1007/s11263-010-0390-2.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL https://arxiv.org/abs/2511.08544.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/97af07a14cacba681feacf3012730892-Abstract.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020. URL https://arxiv.org/abs/2006.07159.

Black Forest Labs. FLUX.2: Frontier visual intelligence. Official model release, 2025. URL https://bfl.ai/blog/flux-2.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining Discriminative Components with Random Forests. In *European Conference on Computer Vision*, 2014. URL https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/.

Alessandro Burzio, Tobias Fischer, Sven Elflein, Qunjie Zhou, Riccardo de Lutio, Jiawei Ren, Jiahui Huang, Shengyu Huang, Marc Pollefeys, Laura Leal-Taixé, Zan Gojcic, and Haithem Turki. Déjà view: Looping transformers for multi-view 3D reconstruction. In *Advances in Neural Information Processing Systems*, 2026. URL https://arxiv.org/abs/2605.30215.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *European Conference on Computer Vision*, 2012. URL https://files.is.tue.mpg.de/black/papers/ButlerECCV2012.pdf.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Mathilde_Caron_Deep_Clustering_for_ECCV_2018_paper.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020a. URL https://proceedings.mlr.press/v119/chen20s.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020b. URL https://proceedings.mlr.press/v119/chen20j.html.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, 2020c. URL https://arxiv.org/abs/2006.10029.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020d. URL https://arxiv.org/abs/2003.04297.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Cimpoi_Describing_Textures_in_2014_CVPR_paper.html.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep Learning for Classical Japanese Literature. *arXiv preprint arXiv:1812.01718*, 2018. URL https://arxiv.org/abs/1812.01718.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/20568692db622456cc42a2e853ca21f8-Abstract.html.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali

Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Deitke_Molmo_and_PixMo_Open_Weights_and_Open_Data_for_State-of-the-Art_CVPR_2025_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009. URL https://doi.org/10.1109/CVPR.2009.5206848.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *International Conference on Computer Vision*, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Doersch_Unsupervised_Visual_Representation_ICCV_2015_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Christoffer Edlund, Timothy R. Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. LIVECell—A Large-Scale Dataset for Label-Free Live Cell Segmentation. *Nature Methods*, 2021. URL https://doi.org/10.1038/s41592-021-01249-6.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Ángel Bautista, Vaishaal Shankar, Alexander T Toshev, Joshua M. Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/el-nouby24a.html.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *Conference on Computer Vision and Pattern Recognition*, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ericsson_How_Well_Do_Self-Supervised_Models_Transfer_CVPR_2021_paper.html.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. VOC2012 challenge website, 2012. URL https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2012/index.html.

Qihang Fan, Huaibo Huang, and Ran He. Breaking the low-rank dilemma of linear attention. In *Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fan_Breaking_the_Low-Rank_Dilemma_of_Linear_Attention_CVPR_2025_paper.html.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G. Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fini_Multimodal_Autoregressive_Pre-training_of_Large_Vision_Encoders_CVPR_2025_paper.html.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision Meets Robotics: The KITTI Dataset. *International Journal of Robotics Research*, 2013. URL https://www.cvlibs.net/datasets/kitti/.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *International Conference on Computer Vision*, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Goyal_Scaling_and_Benchmarking_Self-Supervised_Visual_Representation_Learning_ICCV_2019_paper.html.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. URL https://arxiv.org/abs/2103.01988.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1706.04261.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.html.

Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. D^2USt3R: Enhancing 3D reconstruction for dynamic scenes. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.06264.

Ali Hatamizadeh and Jan Kautz. MambaVision: A hybrid mamba-transformer vision backbone. In *Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hatamizadeh_MambaVision_A_Hybrid_Mamba-Transformer_Vision_Backbone_CVPR_2025_paper.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer.

The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

Ali Hojjat, Janek Haberer, Soren Pirk, and Olaf Landsiedel. ThinkingViT: Matryoshka thinking vision transformer for elastic inference. In *Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hojjat_ThinkingViT_Matryoshka_Thinking_Vision_Transformer_for_Elastic_Inference_CVPR_2026_paper.html.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. URL https://arxiv.org/abs/1704.04861.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. In *Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pFyzqbUiF9.

Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. URL https://arxiv.org/abs/1608.08614.

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv preprint arXiv:1602.07360*, 2016. URL https://arxiv.org/abs/1602.07360.

Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Islam_A_Broad_Study_on_the_Transferability_of_Visual_Representations_With_ICCV_2021_paper.html.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. URL https://arxiv.org/abs/1705.06950.

Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Kendall_PoseNet_A_Convolutional_ICCV_2015_paper.html.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IVAlYCqdvW.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kolesnikov_Revisiting_Self-Supervised_Visual_Representation_Learning_CVPR_2019_paper.html.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013. URL https://openaccess.thecvf.com/content_iccv_workshops_2013/W19/html/Krause_3D_Object_Representations_2013_ICCV_paper.html.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Hildegard Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A Large Video Database for Human Motion Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011. URL https://doi.org/10.1109/ICCV.2011.6126543.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-Level Concept Learning through Probabilistic Program Induction. *Science*, 2015. URL https://doi.org/10.1126/science.aab3050.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998. URL https://doi.org/10.1109/5.726791.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-73220-1_5.

Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Li_CrowdPose_Efficient_Crowded_Scenes_Pose_Estimation_and_a_New_Benchmark_CVPR_2019_paper.html.

Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. MAGE: MAsked generative encoder to unify representation learning and image synthesis. In *Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_MAGE_MAsked_Generative_Encoder_To_Unify_Representation_Learning_and_Image_CVPR_2023_paper.html.

Tom Li and Lawrence Wong. Industrial-iSeg Dataset. Industrial-iSeg dataset website, 2024. URL https://industrial-iseg.com/downloads/.

Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Li_MegaDepth_Learning_Single-View_CVPR_2018_paper.html.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yirunib8l8.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations*, 2014a. URL https://arxiv.org/abs/1312.4400.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014b. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/baa2da9ae4bfed26520bb61d259a3653-Abstract-Conference.html.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Conference on Computer Vision and Pattern Recognition*, 2022a. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Swin_Transformer_V2_Scaling_Up_Capacity_and_Resolution_CVPR_2022_paper.html.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Computer Vision and Pattern Recognition*, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_3DSRBench_A_Comprehensive_3D_Spatial_Reasoning_Benchmark_ICCV_2025_paper.html.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft. *arXiv preprint arXiv:1306.5151*, 2013. URL https://arxiv.org/abs/1306.5151.

Meta AI. Introducing SAM 3D: Powerful 3D reconstruction for physical world images. Official model release, 11 2025. URL https://ai.meta.com/blog/sam-3d/.

Meta AI. SAM 3.1: Release notes. Official model release, 3 2026. URL https://github.com/facebookresearch/sam3/blob/main/RELEASE_SAM3p1.md.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Computer Vision and Pattern Recognition*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Misra_Self-Supervised_Learning_of_Pretext-Invariant_Representations_CVPR_2020_paper.html.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46448-0_32.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Mottaghi_The_Role_of_2014_CVPR_paper.html.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. In *European Conference on Computer Vision*, 2026. URL https://link.springer.com/chapter/10.1007/978-3-032-37369-4_36.

Aditya Murali, Deepak Alapatt, Pietro Mascagni, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. The Endoscapes Dataset for Surgical Scene Segmentation, Object Detection, and Critical View of Safety Assessment: Official Splits and Benchmark. *arXiv preprint arXiv:2312.12429*, 2023. URL https://arxiv.org/abs/2312.12429.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Neuhold_The_Mapillary_Vistas_ICCV_2017_paper.html.

Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. URL https://www.robots.ox.ac.uk/~vgg/data/flowers/102/.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46466-4_5.

Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Noroozi_Boosting_Self-Supervised_Learning_CVPR_2018_paper.html.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

OpenAI. GPT-5.6 Sol. OpenAI API model documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed September 24, 2026.

OpenAI. GPT-6 Astra. OpenAI API model documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed September 24, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and Dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. URL https://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Conference on Computer Vision and Pattern Recognition*, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Pathak_Context_Encoders_Feature_CVPR_2016_paper.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment Matching for Multi-Source Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Peng_Moment_Matching_for_Multi-Source_Domain_Adaptation_ICCV_2019_paper.html.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv preprint arXiv:1704.00675*, 2017. URL https://arxiv.org/abs/1704.00675.

Ariadna Quattoni and Antonio Torralba. Recognizing Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. URL https://groups.csail.mit.edu/vision/torralbalab/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Computer Vision and Pattern Recognition*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Radosavovic_Designing_Network_Design_Spaces_CVPR_2020_paper.html.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representation*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. URL https://doi.org/10.1007/s11263-015-0816-y.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Schops_A_Multi-View_Stereo_CVPR_2017_paper.html.

Risa Shinoda, Nakamasa Inoue, Iro Laina, Christian Rupprecht, and Hirokatsu Kataoka. AnimalClue: Recognizing Animals by their Traces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shinoda_AnimalClue_Recognizing_Animals_by_their_Traces_ICCV_2025_paper.html.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Shotton_Scene_Coordinate_Regression_2013_CVPR_paper.html.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, 2012. URL https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.1556.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026. URL https://arxiv.org/abs/2605.18324.

Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Singh_Revisiting_Weakly_Supervised_Pre-Training_of_Visual_Perception_Models_CVPR_2022_paper.html.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Song_SUN_RGB-D_A_2015_CVPR_paper.html.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012. URL https://arxiv.org/abs/1212.0402.

SpaceNet on Amazon Web Services (AWS). Datasets: The SpaceNet Catalog. SpaceNet dataset repository, 2018. URL https://spacenet.ai/datasets/.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015. URL https://arxiv.org/abs/1505.00387.

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4D video reconstruction with dynamic point maps. In *Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sucar_V-DPM_4D_Video_Reconstruction_with_Dynamic_Point_Maps_CVPR_2026_paper.html.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition*, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller models and faster training. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.mlr.press/v139/tan21a.html.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. MnasNet: Platform-aware neural architecture search for mobile. In *Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Tan_MnasNet_Platform-Aware_Neural_Architecture_Search_for_Mobile_CVPR_2019_paper.html.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, 2020. URL https://link.springer.com/chapter/10.1007/978-3-030-58621-8_45.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.html.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Touvron_Going_Deeper_With_Image_Transformers_ICCV_2021_paper.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *European Conference on Computer Visio*, 2022. URL https://link.springer.com/chapter/10.1007/978-3-031-20053-3_27.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Elias Ramzi, Spyros Gidaris, Andrei Bursuc, and Yuki M Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. In *Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Venkataramanan_Franca_Nested_Matryoshka_Clustering_for_Scalable_Visual_Representation_Learning_CVPR_2026_paper.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21/files/CUB_200_2011.pdf.

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.13549.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. In *Computer Vision and Pattern Recognition*, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_VGGT-ohm_CVPR_2026_paper.html.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_VideoMAE_V2_Scaling_Video_Masked_Autoencoders_With_Dual_Masking_CVPR_2023_paper.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Computer Vision and Pattern Recognition*, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wang_Pyramid_Vision_Transformer_A_Versatile_Backbone_for_Dense_Prediction_Without_ICCV_2021_paper.html.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024b. URL https://link.springer.com/chapter/10.1007/978-3-031-73013-9_23.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=DTQIjngDta.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jerome Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Weinzaepfel_CroCo_v2_Improved_Cross-view_Completion_Pre-training_for_Stereo_Matching_and_ICCV_2023_paper.html.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Woo_ConvNeXt_V2_Co-Designing_and_Scaling_ConvNets_With_Masked_Autoencoders_CVPR_2023_paper.html.

Bing Wu et al. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025. URL https://arxiv.org/abs/2511.18870.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wu_CvT_Introducing_Convolutions_to_Vision_Transformers_ICCV_2021_paper.html.

Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_IP102_A_Large-Scale_Benchmark_Dataset_for_Insect_Pest_Recognition_CVPR_2019_paper.html.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN Database: Large-Scale Scene Recognition from Abbey to Zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010. URL https://vision.princeton.edu/projects/2010/SUN/.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.html.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xie_SimMIM_A_Simple_Framework_for_Masked_Image_Modeling_CVPR_2022_paper.html.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. URL https://arxiv.org/abs/2107.00641.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011. URL https://people.csail.mit.edu/khosla/papers/iccv2011_yao_full.pdf.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yu_BDD100K_A_Diverse_Driving_Dataset_for_Heterogeneous_Multitask_Learning_CVPR_2020_paper.html.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Computer Vision and Pattern Recognition*, 2022. URL Yu_MetaFormer_Is_Actually_What_You_Need_for_Vision_CVPR_2022_paper.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Yuan_Tokens-to-Token_ViT_Training_Vision_Transformers_From_Scratch_on_ImageNet_ICCV_2021_paper.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. URL https://arxiv.org/abs/1910.04867.

Gaifan Zhang and Danushka Bollegala. Map of Encoders – mapping sentence encoders using quantum relative entropy. In *Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://aclanthology.org/2026.acl-long.645/.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lJpqxFgWCM.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision*, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46487-9_40.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhang_Split-Brain_Autoencoders_Unsupervised_CVPR_2017_paper.html.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_ShuffleNet_An_Extremely_CVPR_2018_paper.html.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Computer Vision and Pattern Recognition*, 2017. URL https://doi.org/10.1109/CVPR.2017.544.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. URL https://doi.org/10.1109/TPAMI.2017.2723009.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5APgTKsnx8.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zoph_Learning_Transferable_Architectures_CVPR_2018_paper.html.

APPENDIX

A ADDITIONAL DISCUSSION, LIMITATIONS, AND FUTURE DIRECTIONS

A.1 DISCUSSION

Toward General-Purpose Visual Representations. This study compares VSSL, geometric, generative, and video models from the deep learning era through the lens of reusable visual representations. Although progress extends beyond image classification, strengths depend on the task and evaluation protocol. Here, generality refers to the breadth of transfer achieved by a specific checkpoint and feature-extraction interface under stated conditions, rather than the model's original function or whole-system capabilities.

The DINO family has advanced image-representation transfer (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2026); geometric models have pursued structure estimation and reconstruction (Wang et al., 2025; 2026a); video SSL has developed representation learning from video (Tong et al., 2022; Mur-Labadia et al., 2026); and generative models have explored connections between synthesis and representation learning (Li et al., 2023; Singh et al., 2026). These developments do not form a simple history of one approach replacing another. Whether their distinct strengths can be integrated into a single representation remains an open question.

Remaining Opportunities in Learning Design. In our unified evaluation (Figure 3; Tables 36–37), DINOv3 (Siméoni et al., 2026) demonstrates broad transfer even in comparisons that include VLMs and multimodal world models (MMWMs) as references. However, V-JEPA 2.1 (Mur-Labadia et al., 2026) and SigLIP 2 Giant (Tschannen et al., 2025) outperform DINOv3-7B on SSv2 under unified LP (Table 36). Representations learned from still images are broadly useful, but transfer to video classification alone does not establish an understanding of temporal dynamics.

In our diagnostic experiments (Figure 5; Tables 18–19), the multi-crop configuration of Instance Discrimination (Wu et al., 2018) improves all five tasks over the control configuration within the same experimental series. The shared-head DINOv3 configuration with a broader masking range (Table 15) improves classification, detection, and segmentation, largely preserves depth performance, and slightly reduces video-classification performance. These are separate experiments with smaller backbones, not improvements to DINOv3-7B. The results identify opportunities for further validation in both classical and modern learning designs: there is still room to learn from classical VSSL and advance modern approaches.

Specialized Functions and Representation Reuse. Specialization in a particular function does not guarantee uniform improvements in other uses. Among the evaluated VGGT-Ω (Wang et al., 2026a) configurations (Tables 24–25), aggregator features perform better on image classification, whereas scene features perform better on semantic segmentation under the category-specific protocol. The feature-extraction interface therefore also affects transfer performance.

RAE (Zheng et al., 2026) and RAEv2 (Singh et al., 2026) likewise require distinguishing the encoder, tokenizer, generator, and decoder (Tables 27–28). The evaluated RAEv2-K7 pathway uses a DINOv3-L-based tokenizer without loading the generator or decoder. Its strengths cannot automatically be attributed to generative training. Moreover, differences in data, scale, and architecture mean that these comparisons do not isolate the causal effects of specialization.

Reconstruction, image synthesis, and latent prediction offer connections to representation learning (He et al., 2022; Li et al., 2023; Assran et al., 2023; Bardes et al., 2024). However, this study does not measure generation quality, and its relationship with transfer performance remains untested. Future work should evaluate original functions alongside downstream transfer, distinguishing strengths inherited from an existing encoder from the effects of additional training.

Distinguishing Whole-System Capabilities. Multimodal-system performance may reflect not only visual representations but also language knowledge, additional training, inference-time computation, and tool use. OOur unified comparison, including Cosmos 3 Super (NVIDIA, 2026), evaluates the reuse of specified visual components rather than whole-system capabilities.

The frontier-system reference comparison (Figure 6; Appendix C.15) is also a separate evaluation with matched samples and task formulations, distinct from both the standard Basic Five and com-

parisons with aligned pre-training conditions. The usefulness of internal representations and whole-system performance are complementary but separate evaluation targets.

A.2 LIMITATIONS

Model coverage and scope of control. Representation evaluation primarily covers models whose internal features are accessible. Models differ in pre-training data, scale, resolution, and computational cost, preventing attribution of performance differences to learning objectives alone (Kolesnikov et al., 2019; Goyal et al., 2019). CTRL aligns key conditions while retaining method-specific mechanisms. Equal epoch counts do not imply equal computation, and data-scale comparisons also change training duration.

Dependence on evaluation protocols. Shared conditions enable comparison but do not guarantee each model's best performance. LP includes minimal task-specific predictors and is not necessarily a strictly linear mapping. LP/AP freeze the feature extractor, whereas FT updates it separately for each task. Optimization and augmentation also differ across evaluation tracks, so ranking changes cannot be attributed solely to the predictor or the scope of parameter updates.

Repetitions and evaluation coverage. Repetition counts and completed evaluations vary across results, and learning-design diagnostics use a single seed. The statistical significance of small differences and ranking reversals, as well as the reproducibility of improvements, requires further validation. Because aggregate results depend on valid shared entries and their weighting, we report evaluation coverage, repetition counts, and individual scores alongside summaries.

Limits of the measured capabilities. The Basic Five and extended evaluations do not exhaust visual intelligence. We do not directly evaluate generation quality, long-horizon prediction, active perception, action generation, or planning. Strong performance after independent task-specific FT also does not demonstrate that a single model simultaneously retains multiple capabilities.

Comparison with closed models. Where possible, we conduct downstream training from publicly available checkpoints or reproduce pre-training following the original methods. However, closed models may achieve higher scores, and our inability to compare them comprehensively is a limitation. Although it does not establish an overall ranking of visual capabilities, the reference comparison with the frontier model GPT-6 Astra (OpenAI, 2026b) offers a partial view of this gap and indicates that vision-native models still have room for improvement.

A.3 FUTURE DIRECTIONS

One promising direction is to learn geometric, generative, and spatiotemporal structure while preserving the usefulness of existing representations. For both joint pre-training and continual learning, evaluating corresponding checkpoints before and after training on newly acquired functions and transfer to existing tasks would directly test capability acquisition and retention. It is important to control data, architecture, feature extraction, and training budgets where feasible, and to distinguish a single shared representation from separately task-adapted models.

Connections to interaction, action, and planning are also important research directions. However, action-conditioned learning and related approaches extend beyond this paper's strict vision-native scope, requiring explicit identification of the additional learning signals (Assran et al., 2025). A central question is whether capabilities developed in different fields can be connected within one representation that broadens its uses while preserving existing strengths. Our experimental records and evaluation framework provide a starting point for this investigation.

B MODELS, TRAINING, AND EVALUATION SETTINGS

This appendix identifies the evaluated models, pretraining supervision, and experiment-specific settings. Sections 3, 4, and 5 cover historical/category-specific comparisons, unified representative-model evaluation, and controlled pretraining/design changes, respectively. Scores are in Appendix C; task definitions and shared evaluation protocols are in Appendix E.

Checkpoint-level classification. *Strict vision-native* pretraining uses images, video, or visual/geometric observations without semantic category labels, text or language-derived targets, audio, or actions as direct inputs or targets. Depth, poses, point maps, and motion are admissible geometric targets. *Reference* covers supervised, vision–language, class-conditioned, and multimodal comparisons; *boundary vision-native* marks references that add language alignment or action conditioning to a visual model. Status applies to the evaluated checkpoint and component. Removing conditioning modules at evaluation does not erase pretraining supervision, and distilled models retain their teachers' supervision provenance. Downstream labels do not change pretraining status.

Reading the catalog. Entries distinguish published objectives from evaluated checkpoints and feature interfaces. *Component-dependent* means the component or training stage is not sufficiently identified to confirm strict vision-native status. One family can cover multiple scored configurations; the architecture comparison contains 48. ASIS includes public releases and local reproductions; CTRL backbones are trained locally. Checkpoint filenames are model-specific, and release dates are not execution dates. Shared settings are stated once below; model-specific exceptions remain in the catalog.

B.1 VISUAL SELF-SUPERVISED LEARNING

The primary comparison contains 32 methods: 31 selected image methods below and the V-JEPA 1 ViT-H ASIS checkpoint in Table 6. DeepCluster, CPC, BEiT v1, and SimMIM are additional ASIS methods outside CTRL; Table 10 also retains the unscored NEPA entry. BEiT v1 uses the DALL-E dVAE tokenizer, not BEiT v2's VQ-KD targets; SimCLR v2 is evaluated before semi-supervised fine-tuning or distillation.

CTRL trains ViT-B/16 on unlabeled ImageNet-1K for 300 epochs using seed 42, 224-pixel inputs, batch 1,024, AdamW at 6×10^{-4}, weight decay 0.05, 10 warmup epochs, and cosine decay. Essential masking, cropping, teacher, queue, and objective mechanisms are retained. Results at epochs 100/200/300 come from the same trajectory, not independent repetitions. Method-specific probes remain distinct: equal epochs do not imply equal compute or identical downstream optimization (Appendix E.2).

Table 1: Visual Self-Supervised Learning: models, supervision status, and feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
VAE (Kingma & Welling, 2014); 2013-12	Strict vision-native	ASIS trains a local convolutional VAE and reads the reconstruction bottleneck. Under shared settings the objective remains reconstruction with a KL term; the CNN is replaced by the shared ViT-B/16. Downstream adapters convert the latent code to the required global or spatial interface. Local ASIS record: 300 epochs, `checkpoint_epoch_300.pth`.
Context Prediction (Doersch et al., 2015); 2015-05	Strict vision-native	ASIS follows an official-style AlexNet spatial-context pretext and probes fc6 after spatial pooling. Local ASIS record: 1M steps, `final.pth`.
Colorization (Zhang et al., 2016); 2016-03	Strict vision-native	ASIS trains a local CNN for cross-channel colour prediction and uses GAP features. Under shared settings the Lab colour target is kept and colour jitter is omitted. Downstream tasks apply the Lab conversion inside the frozen adapter. Local ASIS record: `best_model.pth`.

Continued on the next page

Table 1 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
Context Encoder (Pathak et al., 2016); 2016-04	Strict vision-native	ASIS does not rerun inpainting: official Caffe features are converted to PyTorch and probed. Under shared settings a local inpainting-style objective is trained on the shared ViT-B/16 recipe.
Jigsaw Puzzles (Noroozi & Favaro, 2016); 2016-03	Strict vision-native	ASIS trains a local AlexNet CFN on 1,000 tile permutations. The ImageNet probe uses 224-pixel full-image crops and 4×4 CFN pooling. Local ASIS record: 300 epochs, `best_model.pth`.
Rotation Prediction (Gidaris et al., 2018); 2018-03	Strict vision-native	ASIS trains official-style AlexNet-BN on four-way rotation and probes conv5 GAP. Local ASIS record: 50 epochs, `checkpoint_epoch_50.pth`.
SplitBrain (Zhang et al., 2017); 2016-11	Strict vision-native	ASIS trains a local AlexNet for cross-channel prediction. Under shared settings the split-channel target is kept, with Lab conversion inside the frozen adapter. Local ASIS record: 99 epochs, `checkpoint_epoch_99.pth`.
Jigsaw Puzzle++ (Noroozi et al., 2018); 2018-05	Strict vision-native	ASIS trains a local AlexNet with transformation-generated cluster targets.
Instance Discrimination (Wu et al., 2018); 2018-05	Strict vision-native	ASIS trains ResNet-50 with a memory-bank instance objective and probes GAP features. Under shared settings the memory bank is retained. The CTRL comparison (Section 5.1) is distinct from the design changes in Section 5.2. Local ASIS record: 200 epochs, `checkpoint_epoch_200.pth`.
CMC (Tian et al., 2020); 2019-06	Strict vision-native	ASIS trains a two-stream AlexNet on Lab views and reads layer-5 features. Under shared settings the cross-view contrastive objective is kept, with Lab conversion in the adapter. Local ASIS record: 200 epochs, `checkpoint_epoch_0200.pth`.
MoCo v1 (He et al., 2020); 2019-11	Strict vision-native	ASIS retrains ResNet-50 with a 65,536-slot queue, momentum 0.999, temperature 0.07, shuffle BN, and step learning-rate decay. The ImageNet probe is a cached cosine linear eval, not official `main_lincls.py`. Under shared settings the queue and momentum are retained on the shared AdamW recipe. Local ASIS record: 200 epochs, `checkpoint_epoch_200.pth`.
PIRL (Misra & van der Maaten, 2020); 2019-12	Strict vision-native	ASIS trains ResNet-50 for pretext-invariant instance learning. Local ASIS record: 800 epochs, `checkpoint_epoch_800.pth`.
SimCLR v1 (Chen et al., 2020b); 2020-02	Strict vision-native	ASIS trains ResNet-50 with two-view NT-Xent. Projection heads are discarded; the frozen encoder is probed. Local ASIS record: 1000 epochs, `checkpoint_epoch_1000.pth`.
MoCo v2 (Chen et al., 2020d); 2020-03	Strict vision-native	ASIS retrains ResNet-50 with temperature 0.2, an MLP projection, and shuffle BN. The ImageNet probe is not a byte-identical official linear schedule. Under shared settings those contrastive extras are kept. Local ASIS record: 200 epochs, `checkpoint_epoch_200.pth`.
SimCLR v2 (Chen et al., 2020c); 2020-06	Strict vision-native	ASIS uses the local ResNet-50 SSL checkpoint at epoch 800, before semi-supervised fine-tune or distillation. Under shared settings only the contrastive loss is trained from scratch on the shared recipe. Local ASIS record: 800 epochs, `checkpoint_epoch_800.pth`.
SwAV (Caron et al., 2020); 2020-06	Strict vision-native	ASIS: locally trained ResNet-50 with online clustering and multi-crop. CTRL retains multi-crop and swapped-prediction prototypes. Its COCO detector uses learning rate 5×10^{-4} rather than the shared 5×10^{-3}; the exception is retained in protocol comparisons. Local ASIS record: 200 epochs, `checkpoint_epoch_200.pth`.

Continued on the next page

Table 1 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
SeLa (Asano et al., 2020); 2019-11	Strict vision-native	ASIS trains ResNetV2 with self-labels from visual clustering. Local ASIS record: 400 epochs, `checkpoint_epoch_400.pth`.
BYOL (Grill et al., 2020); 2020-06	Strict vision-native	ASIS trains ResNet-50 with a momentum teacher and a predictor, without negatives. Evaluation uses the specified student or target encoder, not a tuned ensemble. Local ASIS record: 1000 epochs, `checkpoint_epoch_1000.pth`.
SimSiam (Chen & He, 2021); 2020-11	Strict vision-native	ASIS: locally trained ResNet-50 with stop-gradient Siamese learning and SyncBN; ImageNet uses a 90-epoch LARS linear classifier. CTRL retains the predictor and stop-gradient under the shared pretraining recipe. Local ASIS record: 100 epochs, `checkpoint_epoch_100.pth`.
Barlow Twins (Zbontar et al., 2021); 2021-03	Strict vision-native	ASIS uses the BF16 epoch-1000 ResNet-50 redundancy-reduction run. Local ASIS record: 1000 epochs, `checkpoint_epoch_1000.pth`.
MoCo v3 (Chen et al., 2021); 2021-04	Strict vision-native	ASIS trains ViT-B/16 for 300 epochs at batch 4096, AdamW learning rate 2.4×10^{-3}, weight decay 0.1, 40-epoch warmup, momentum $0.99\rightarrow1.0$, and temperature 0.2, then a 90-epoch DDP linear probe. Under shared settings τ=0.2 and m=0.99 are kept, but batch, learning rate, weight decay, and warmup follow the shared recipe rather than the ASIS scale. Local ASIS record: 300 epochs, `checkpoint_epoch_300.pth`.
DINO (Caron et al., 2021); 2021-04	Strict vision-native	ASIS trains ViT-S/16 and probes teacher last-four CLS tokens. The CTRL ImageNet probe remains DINO-style, not Section 4 LINEAR. Local ASIS record: 100 epochs, `checkpoint_epoch_100.pth`.
MAE (He et al., 2022); 2021-11	Strict vision-native	ASIS trains ViT-L/16 for 1,600 epochs with 75% masking and pixel reconstruction; evaluation uses the encoder CLS. Under shared settings the same mask ratio and `norm_pix_loss` are kept, colour jitter is omitted, and the decoder is discarded after pretraining. Local ASIS record: 1600 epochs, `checkpoint_epoch_1600.pth`.
iBOT (Zhou et al., 2022); 2021-11	Strict vision-native	ASIS trains ViT-S/16 and probes teacher last-four CLS tokens with distributed training. Local ASIS record: 800 epochs, `checkpoint_epoch_800.pth`.
MSN (Assran et al., 2022); 2022-04	Strict vision-native	ASIS trains ViT-S/16 for masked Siamese learning. Evaluation reads the encoder, not a new prototype classifier. Local ASIS record: 800 epochs, `step1_msn_vits16-latest.pth.tar`.
DINOv2 (Oquab et al., 2024); 2023-04	Strict vision-native	ASIS loads official `dinov2_vitg14` and applies a local cached-feature AdamW linear probe, not the official released linear head. Under shared settings a local ViT-B/16 is trained from scratch and is not the official G/14 file.
I-JEPA (Assran et al., 2023); 2023-01	Strict vision-native	ASIS trains ViT-H/14 for joint-embedding block prediction. Local ASIS record: 300 epochs, `checkpoint_epoch_300.pth`.
AIM (El-Nouby et al., 2024); 2024-01	Strict vision-native	ASIS uses public `apple/aim-600M` with caption-free patch prediction. The selected ASIS ImageNet result uses the official best-layer attentive probe, not the alternative mean-pooled linear probe. CTRL trains the objective locally on ViT-B/16; its result is not a probe-only comparison against ASIS.
Franca (Venkataramanan et al., 2026); 2025-07	Strict vision-native	ASIS loads official ViT-B/14 ImageNet-21K weights and probes the CLS token. Under shared settings a local ViT-B/16 is trained on unlabelled ImageNet-1K and is not the official In21K run.

Continued on the next page

Table 1 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
DINOv3 (Siméoni et al., 2026); 2025-08	Strict vision-native	ASIS loads official `dinov3-vitb16-pretrain-lvd1689m` at 512×512 with a local cached-feature probe. Under shared settings a local ViT-B/16 is trained and EMA teacher weights are preferred at evaluation. This is not the Section 4 DINOv3-7B checkpoint.
LeJEPA (Balestriero & LeCun, 2025); 2025-11	Strict vision-native	ASIS uses local ViT-B/16 with the LeJEPA objective. CTRL uses the shared optimizer and schedule. Its recorded COCO Faster R-CNN head re-evaluation uses learning rate 5×10^{-4} rather than the shared 5×10^{-3}. Evaluation reads the frozen encoder. Local ASIS record: 100 epochs, `checkpoint_epoch_100.pth`.
Additional ASIS methods (outside CTRL)		
DeepCluster (Caron et al., 2018); 2018-07	Strict vision-native	Alternates clustering of image features and learning from the resulting assignments. The additional ASIS entry is not a locally trained ViT-B/16 CTRL trajectory.
CPC (van den Oord et al., 2018); 2018-07	Strict vision-native	Contrastive prediction of latent representations with an InfoNCE objective. The catalog concerns the visual ASIS configuration, not the speech or language applications of the framework.
BEiT v1 (Bao et al., 2022); 2021-06	Strict vision-native	Masked prediction of discrete visual tokens from the DALL-E dVAE. This additional ASIS entry is distinct from BEiT v2 and outside the controlled comparison.
SimMIM (Xie et al., 2022); 2021-11	Strict vision-native	Masked image modeling with direct pixel prediction. Included in the additional ASIS results, outside the selected 32-method controlled comparison.

B.2 Supervised Models and Architecture References

Table 23 contains 48 architecture and training configurations. Most are supervised references; SEER is self-supervised, SWAG uses hashtag supervision, and ConvNeXt V2 here is supervised fine-tuning after FCMAE. NIN and Highway use CIFAR-10 weights. Variant-specific identifiers remain attached to each task result rather than being collapsed into one family checkpoint.

Table 2: Architecture References: models, supervision status, and feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
LeNet-5 (LeCun et al., 1998); 1998	Reference	Local ImageNet-supervised weights. Global features come from the final fully connected stack after the 32×32-style interface is adapted to the Basic Five loaders. Local ImageNet-1K training: 10 epochs.
AlexNet (Krizhevsky et al., 2012); 2012	Reference	ImageNet-supervised AlexNet. ImageNet uses the published 1,000-way classifier; dense and video tasks use the last convolutional map.
Network in Network (Lin et al., 2014a); 2013-12	Reference	CIFAR-10 supervised `nin_cifar10`, not ImageNet weights. Inputs are resized to the CIFAR-trained interface; scores are reference supervised transfer. Locally trained for 308 epochs.
VGG (Simonyan & Zisserman, 2015); 2014-09	Reference	Official VGG-16 ImageNet-supervised weights. Features come from the final convolutional stack before the classifier. Weight identifier: `vgg16-397923af.pth`.
GoogLeNet (Szegedy et al., 2015); 2014-09	Reference	`GoogLeNet_Weights.IMAGENET1K_V1`. Auxiliary classifiers are unused. The main-branch pool5 representation is used.

Continued on the next page

Table 2 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
Highway Network (Srivastava et al., 2015); 2015-05	Reference	CIFAR-10 supervised Highway-19 (`highway_cifar19`), not ImageNet pretraining. The same CIFAR-trained caveat as Network in Network applies. Locally trained for 200 epochs.
ResNet-50 / ResNet-152 (He et al., 2016); 2015-12	Reference	The ResNet-50 configuration uses `ResNet50_Weights.IMAGENET1K_V1`. ImageNet uses the published classifier; dense and video tasks read the final residual map. ResNet-152 is a separately scored configuration, not a replacement for ResNet-50 in selected columns.
SqueezeNet (Iandola et al., 2016); 2016-02	Reference	`SqueezeNet1_0_Weights.IMAGENET1K_V1`. Fire-module features are globally pooled or reshaped; no local SqueezeNet pretraining was run.
Xception (Chollet, 2017); 2016-10	Reference	timm/Cadene ImageNet-supervised Xception. Entry-flow and exit-flow remain frozen. Evaluation uses the 299-aware preprocessing of that checkpoint. Weight identifier: `xception-43020ad28.pth`.
DenseNet-121 / DenseNet-161 (Huang et al., 2017); 2016-08	Reference	DenseNet-121 is trained locally on ImageNet-1K for 90 epochs; features come from the final dense block. DenseNet-161 is a separately scored configuration.
ResNeXt (Xie et al., 2017); 2016-11	Reference	ResNeXt-50 32×4d ImageNet-supervised weights. Cardinality groups stay frozen; heads read the final 2048-D map.
MobileNet (Howard et al., 2017); 2017-04	Reference	`google/mobilenet_v1_1.0_224`. Depthwise-separable features are frozen. ImageNet uses the published 224-pixel classifier.
ShuffleNet (Zhang et al., 2018); 2017-07	Reference	Locally trained ShuffleNetV1 on ImageNet-1K for 100 epochs. Channel-shuffle stage features are used for transfer.
NASNet (Zoph et al., 2018); 2017-07	Reference	NASNet-A Large ImageNet-supervised weights. Cell outputs stay frozen; the published stem is unchanged. Weight identifier: `nasnetalarge.tf_in1k`.
SENet / SENet-154 (Hu et al., 2018); 2017-09	Reference	The SENet-154 configuration uses ImageNet-supervised `senet154` weights. Squeeze-and-excitation gates remain part of the frozen backbone. The SENet and explicitly named SENet-154 score rows retain their separate recorded identities.
MobileNetV2 (Sandler et al., 2018); 2018-01	Reference	Official ImageNet-supervised MobileNetV2. Inverted residuals stay frozen under the published 224-pixel normalization. Weight identifier: `IMAGENET1K_V1`.
MnasNet (Tan et al., 2019); 2018-07	Reference	MnasNet1.0 ImageNet-supervised weights. The platform-search architecture is used as released; we did not rerun NAS. Weight identifier: `IMAGENET1K_V1`.
EfficientNet (Tan & Le, 2019); 2019-05	Reference	EfficientNet-B0 ImageNet-supervised weights. Compound-scaled MBConv features are read at the final stage. Weight identifier: `IMAGENET1K_V1`.
RegNet (Radosavovic et al., 2020); 2020-03	Reference	RegNetY-800MF ImageNet-supervised weights. The AnyNet-parameterized body is frozen; this row is not SEER. Weight identifier: `IMAGENET1K_V1`.
ViT-B/16, L/16, H/14 (Dosovitskiy et al., 2021); 2020-10	Reference	Supervised ViT checkpoints. ViT-B/16, ViT-L/16, and ViT-H/14 are separately listed score configurations. Global and dense interfaces retain the checkpoint's CLS and patch features, respectively.
T2T-ViT (Yuan et al., 2021); 2021-01	Reference	Official T2T-ViT-14 ImageNet-supervised checkpoint. Tokens-to-token embeddings stay frozen and supply CLS and spatial tokens. Weight identifier: `81.5_T2T_ViT_14.pth.tar`.

Continued on the next page

Table 2 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
CaiT (Touvron et al., 2021); 2021-03	Reference	Official CaiT-S24 ImageNet-supervised checkpoint. Class-attention tokens are used for global tasks; patch tokens for dense tasks.
PVT (Wang et al., 2021); 2021-02	Reference	Official PVT-Small ImageNet-supervised checkpoint. Pyramid stage maps feed detection and segmentation; the final stage is pooled for classification. Weight identifier: `pvt_small.pth`.
CvT (Wu et al., 2021); 2021-03	Reference	`microsoft/cvt-13`. Convolutional token embeddings remain frozen. No local CvT pretraining was run.
Swin Transformer / Swin-B (Liu et al., 2021); 2021-03	Reference	Supervised shifted-window representations. The explicitly named Swin-B configuration uses official ImageNet-supervised weights (`IMAGENET1K_V1`); it remains separate from the generic Swin Transformer score row. Dense heads read spatial feature maps.
Focal Transformer (Yang et al., 2021); 2021-07	Reference	Focal-Tiny, ImageNet-supervised checkpoint. Focal self-attention stages provide frozen features.
MLP-Mixer (Tolstikhin et al., 2021); 2021-05	Reference	Mixer-B/16 trained on ImageNet-21K then fine-tuned on ImageNet-1K. Channel- and token-mixing blocks are frozen.
PoolFormer (Yu et al., 2022); 2021-11	Reference	PoolFormer-S12 ImageNet-supervised checkpoint. Pooling-as-token-mixer features replace attention and stay frozen.
CoAtNet (Dai et al., 2021); 2021-06	Reference	CoAtNet-0 ImageNet-supervised checkpoint. Convolution–attention stages remain frozen. Weight identifier: `coatnet_0_rw_224.sw_in1k`.
EfficientNetV2 (Tan & Le, 2021); 2021-04	Reference	Official EfficientNetV2-S supervised checkpoint. Fused-MBConv features are frozen; training-time progressive resize is not reapplied at evaluation. Weight identifier: `IMAGENET1K_V1`.
SEER RegNetY (Goyal et al., 2021); 2021-03	Strict vision-native	Official `regnety_320.seer` self-supervised Instagram checkpoint. This is the only strict vision-native architecture row. No hashtag or category label is used at the evaluated stage.
ConvNeXt / ConvNeXt-L (Liu et al., 2022b); 2022-01	Reference	Supervised modernized ConvNet representations. The explicitly named ConvNeXt-L configuration uses official ImageNet-supervised weights (`IMAGENET1K_V1`); it remains separate from the generic ConvNeXt score row.
MaxViT (Tu et al., 2022); 2022-04	Reference	MaxViT-Tiny ImageNet-supervised checkpoint. Multi-axis attention blocks remain frozen. Weight identifier: `maxvit_tiny_tf_224.in1k`.
Swin Transformer V2 / Swin V2-B (Liu et al., 2022a); 2021-11	Reference	Supervised shifted-window representations with residual post-normalization and scaled cosine attention. Swin V2-B (`IMAGENET1K_V1`) and the generic Swin Transformer V2 row are scored separately.
SWAG (Singh et al., 2022); 2022-01	Reference	`IMAGENET1K_SWAG_LINEAR_V1`. Instagram hashtag pretraining keeps the row weakly supervised even though ImageNet evaluation is a linear head.
ConvNeXt V2 (Woo et al., 2023); 2023-01	Reference	`convnextv2_tiny_1k_224_ema.pt` is the ImageNet-1K fine-tune after FCMAE. The FCMAE-only checkpoint is not this row.
VMamba (Liu et al., 2024); 2024-01	Reference	Official VMamba-T ImageNet-supervised checkpoint. ImageNet, ADE20K, NYUv2, and SSv2 follow the shared frozen protocol.
MambaVision (Hatamizadeh & Kautz, 2025); 2024-07	Reference	`nvidia/MambaVision-T-1K`. Hybrid Mamba–Transformer stages remain frozen; no local Mamba pretraining was run.

Continued on the next page

Table 2 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
RAVLT (workbook: RA-VLT) (Fan et al., 2025); 2024-11	Reference	Workbook label RA-VLT. Evaluated checkpoint: ImageNet-1K supervised RAVLT-S from `aldjalkdf/RAVLT`. The cited paper is the RALA/RAVLT linear-attention series. No self-supervised RAVLT pretraining is claimed. Weight identifier: `RAVLT_S.pth`.
ThinkingViT (Hojjat et al., 2026); 2025-07	Reference	`thinkingvit_deit-3h-6h-imagenet1k` supervised checkpoint. Matryoshka early-exit branches are not selected per task; the released supervised trunk is frozen.

B.3 3D AND 4D GEOMETRIC MODELS

Static 3D and dynamic 4D models are cataloged separately, matching Tables 24 and 26. VGGT-Ω component ablations follow the 3D comparison in Table 25.

The category-specific CroCo v2, DUSt3R, VGGT, and VGGT-Ω evaluations use native square resolutions of 224/512/518/512 pixels, respectively, with frozen encoder/aggregator tokens and no 3D prediction heads. CroCo uses ImageNet normalization, DUSt3R mean/std 0.5, and VGGT unit-interval RGB with internal normalization. Cached-feature classification uses SGD (momentum 0.9, no weight decay, batch 1,024), cosine decay with 10-epoch warmup, and a learning-rate sweep from 10^{-4} to 10^{-1}, reporting the best validation top-1: 100 head epochs for ImageNet and 50 for eight-frame SSv2. ADE20K uses a 1×1 head (SGD 10^{-2}, 50 epochs); NYUv2 uses a log-depth/SiLog head (AdamW 10^{-3}, 50 epochs; valid range 10^{-3}–10 m). COCO uses a 256-channel single-scale adapter and Faster R-CNN (AdamW 10^{-3}, 12 epochs, batch 2/GPU). These category-specific settings are distinct from the unified 224-pixel representative-model protocol.

Table 3: 3D Geometric Models: models, supervision status, and feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
CroCo v2 (Weinzaepfel et al., 2023); 2022-11	Strict vision-native	Encoder of the official `CroCo_V2_ViTLarge_BaseDecoder.pth`; the decoder is dropped. Single images are encoded to patch tokens. Global tasks mean-pool encoder tokens; dense tasks reshape them to a grid.
DUSt3R encoder (Wang et al., 2024a); 2023-12	Strict vision-native	Encoder of the official `DUSt3R_ViTLarge_BaseDecoder_512_dpt.pth`. Pairwise point-map training is not rerun. Evaluation uses encoder tokens from a single view unless a task supplies a second image.
MASt3R encoder (Leroy et al., 2024); 2024-06	Strict vision-native	Encoder of `naver/MASt3R_ViTLarge_BaseDecoder_512_catmlpdpt_metric`. Matching heads are unused. Features are encoder tokens after the published visual-matching pretraining.
Depth Anything 3 (DA3) (Lin et al., 2026); 2025-11	Strict vision-native	`depth-anything/DA3-LARGE-1.1`. Visual inputs and depth/ray targets define the checkpoint. Evaluation reads the visual encoder, not a complete multi-view reconstruction pipeline.
Pi^3 (Wang et al., 2026b); 2025-07	Strict vision-native	The Pi^3 configuration supplies frozen geometry-trained visual features for Basic Five. Evaluation concerns representation transfer rather than full point-cloud reconstruction.
VGGT (Wang et al., 2025); 2025-03	Strict vision-native	`facebook/VGGT-1B`. Last aggregator patch tokens are used; camera, depth, point-map, and track heads are not the reported representation. Single images use S=1; video uses ordered frames in one aggregator pass.

Continued on the next page

Table 3 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
VGGT-Ω aggregator (Wang et al., 2026a); 2026-05	Strict vision-native	Public non-text `vggt_omega_1b_512.pt`. Final aggregator patch tokens exclude camera/register tokens. Global tasks pool patches; dense tasks retain the grid. Native-resolution category results and unified 224-pixel results are separate. Scene/camera/dense ablations use the same weight file.
VGGT-Ω text-aligned aggregator (Wang et al., 2026a); 2026-05	Boundary vision-native	Public `vggt_omega_1b_256_text.pt`: geometric pretraining followed by language alignment. Evaluated separately from the non-text aggregator and not used as the strict vision-native representative.
VGGT-Ω scene representation (Wang et al., 2026a); 2026-05	Strict vision-native	Scene tokens (camera and register) from the same 512-pixel checkpoint.
VGGT-Ω camera representation (Wang et al., 2026a); 2026-05	Strict vision-native	Camera-head trunk from the same 512-pixel checkpoint.
VGGT-Ω dense representation (Wang et al., 2026a); 2026-05	Strict vision-native	Dense/DPT features from the same 512-pixel checkpoint.

Table 4: 4D Geometric Models: models, supervision status, and feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
Déjà View (Burzio et al., 2026); 2026-05	Strict vision-native	Implementation alias DVLT; `nvidia/dvlt`. Recurrent multi-view reconstruction weights are frozen. Basic Five uses the visual encoder/aggregator pathway, not a full looped reconstruction rollout.
MonST3R (Zhang et al., 2025); 2024-10	Strict vision-native	`Junyi42/MonST3R_PO-TA-S-W_ViTLarge_BaseDecoder_512_dpt`. Encoder tokens only. The patch embedding and all 24 encoder blocks are bit-identical to those of the DUSt3R encoder in Table 3; only the final encoder LayerNorm differs. Dynamic-scene adaptation resides in the unused decoder and heads, so these scores largely reflect the DUSt3R encoder evaluated at 224 pixels. Image tasks encode a single frame; SSv2 encodes sampled frames.
D^{2}USt3R (Han et al., 2025); 2025-04	Strict vision-native. As with MonST3R, the encoder is bit-identical to the DUSt3R encoder except for the final LayerNorm. Time-varying point maps are not decoded, so these scores do not reflect dynamic-scene adaptation.	
V-DPM (Sucar et al., 2026); 2026-01	Strict vision-native	Workbook label VDPM. Checkpoint `edgarsucar/vdpm` (`model.pt`), a VGGT aggregator fine-tuned for dynamic point maps. Evaluation uses the fine-tuned aggregator tokens, not an unfinetuned VGGT.

Continued on the next page

Table 4 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
StreamVGGT (Zhuo et al., 2026); 2025-07	Strict vision-native	`facebook/StreamVGGT-1B`. Streaming visual-geometry tokens are frozen. Causal streaming state is reset per clip or image; no online long-horizon state is carried across datasets.

B.4 Visual Generative Models and Components

Table 27 follows the visual-generative category, including autoregressive and masked-prediction models. Encoder-only transfer and generator-component analyses are distinct experiments. The selected RAE (v1) result uses a DINOv2-B encoder with registers; the selected RAEv2 overview result instead reads generator block 29. Neither is interchangeable with the encoder-only RAEv2 representative. Table 28 reports the remaining component and intervention configurations without repeating the selected overview rows. We evaluate feature transfer, not generation quality; random-weight controls do not demonstrate learned representations.

Table 5: Visual Generative Models and Components: models, supervision status, and feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
iGPT-S (Chen et al., 2020a)	Strict vision-native	Autoregressive pixel prediction. The recorded configuration uses ImageNet-1K at 32×32 pixels; it is not a 224-pixel ViT checkpoint.
MAGE ViT-B/16 encoder (Li et al., 2023); 2022-11	Strict vision-native	Official `mage-vitb-1600ep.pth`, trained on ImageNet-1K for 1,600 epochs. The 256-pixel encoder supplies final-normalized CLS features globally and patch features spatially. The VQGAN and token-decoder paths are not the reported representation.
DiT-XL/2 (Peebles & Xie, 2023); 2022-12	Reference	Official `DiT-XL-2-256x256.pt`, a class-conditioned ImageNet generator. Block features are read at t=0 with fixed class index y=0; this is not the null-class y=1000 used by the RAE component study. Class supervision in pretraining keeps the row a reference.
JiT-L (random weights)	Random control	The evaluated implementation is a randomly initialized architectural placeholder, without pretrained JiT weights. It is an untrained feature control, not a pretrained generative-model result.
data2vec 2.0 (Baevski et al., 2023)	Strict vision-native	The recorded visual configuration uses ImageNet-1K. Its inclusion concerns image-representation transfer, not evaluation of speech or language components.
AIMv2-Large (Fini et al., 2025)	Reference	The recorded Large configuration uses LAION and CC12M. Its multimodal autoregressive pretraining is distinct from the caption-free AIM configuration in the historical VSSL comparison.
RAEv2 K=7 tokenizer (Singh et al., 2026); 2026-05	Strict vision-native	Public DINOv3 ViT-L/16, `dinov3_vitl16_pretrain_lvd1689m`. The $K = 7$ tokenizer averages patch features from layers $\{11, 13, 15, 17, 19, 21, 23\}$ and adds the broadcast mean of the last selected layer. CLS/registers are omitted; global tasks use spatial pooling. The representative-model path loads neither decoder nor DiT and is separate from the component study.

Continued on the next page

Table 5 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
RAE ViT-B encoder / tokenizer (Zheng et al., 2026); 2025-10	Strict vision-native	DINOv2-B with registers supplies the selected encoder-only result and additional tokenizer controls. Encoder extraction removes CLS/register tokens and pools spatial patches. Generator-input features additionally include the latent normalization used by the generator. These interfaces are not interchangeable.
RAE generator / decoder (Zheng et al., 2026); 2025-10	Reference	Class-conditioned generator component study: blocks 0, 6, 13, 20, 27, and 29 at $t \in \{0, 0.1, 0.5, 0.9\}$; null-class baseline $y = 1000$. Decoder features are read at layer 13 and pre-output. Noised-, zero-, and shuffled-latent controls and a random-generator control separate input information from learned transformation.
RAEv2 generator / decoder (Singh et al., 2026); 2026-05	Reference	The DINOv3-L K=7 latent feeds the class-conditioned DiT^{DH}-XL generator. The selected overview row is block 29 at t=0.1, y=1000. The component study separately varies blocks 0/13/20/29 and $t \in \{0, 0.1, 0.5, 0.9\}$, with decoder layer-13/pre-output and input/random-generator controls. Null-class extraction does not change pretraining status.

B.5 Video Self-Supervised Models

Tables 29 contain nine category-specific LP and six attentive configurations, respectively. V-JEPA 1 ViT-H is a separate historical ASIS release. Model size, weight key, temporal sampling, and pretraining stage identify each configuration; action-conditioned weights remain references.

Table 6: Video Self-Supervised Models: models, supervision status, and feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
Shuffle and Learn, ResNet-50 (Misra et al., 2016); 2016-03	Strict vision-native	Evaluated ResNet-50 checkpoint trained with temporal-order verification. Image tasks use a single-frame embedding; SSv2 samples frames and averages frozen features. No optical-flow or action label is used at the evaluated stage.
VideoMAE, ViT-B (Tong et al., 2022); 2022-03	Strict vision-native	`MCG-NJU/videomae-base`. Masked video reconstruction encoder. Image tasks encode a single frame; SSv2 uses uniformly sampled frames. The decoder is dropped.
VideoMAE V2 (Wang et al., 2023)	Component-dependent	Dual-masking video autoencoder, scored separately from VideoMAE in Table 29. The recorded result does not identify whether the checkpoint precedes supervised post-pretraining; it is therefore not assigned confirmed strict vision-native status.
V-JEPA 1, ViT-L (Bardes et al., 2024); 2024-04	Strict vision-native	Public `nielsr/vit-large-patch16-v-jepa`, ViT-L feature path for the category-specific video comparison. Latent prediction without action conditioning; image and video tasks retain their recorded sampling and pooling. Distinct from the ViT-H ASIS release in the next row.
V-JEPA 1, ViT-H (ASIS) (Bardes et al., 2024); 2024-04	Strict vision-native	Official VideoMix2M `vith16.pth.tar`, one of the 32 selected VSSL methods. This public video-pretrained ASIS checkpoint is separate from both the ViT-L video-table path and the locally trained ViT-B/16 CTRL implementation.
V-JEPA 2, ViT-H (Assran et al., 2025); 2025-06	Strict vision-native	`facebook/vjepa2-vith-fpc64-256`, action-free ViT-H image/video backbone. Features come from the target encoder. Actions are not loaded.

Continued on the next page

Table 6 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
V-JEPA 2-AC, ViT-g/16 (Assran et al., 2025); 2025-06	Boundary vision-native	Downloaded `vjepa2_ac_vitg_256.pt`, the action-conditioned stage. The local filename denotes a public weight artifact, not local pretraining. The 256-pixel checkpoint is a reference distinct from action-free V-JEPA 2/2.1.
V-JEPA 2.1, ViT-L (Mur-Labadia et al., 2026); 2026-03	Strict vision-native	Public distilled student `vjepa2_1_vitl_dist_vitG_384.pt`, key `ema_encoder`. Vision-only distillation from ViT-G. This category-specific configuration is not the ViT-g representative.
V-JEPA 2.1, ViT-g (Mur-Labadia et al., 2026); 2026-03	Strict vision-native	Public action-free `vjepa2_1_vitg_384.pt`, key `target_encoder`. The representative uses the final hierarchical normalization over patch tokens. Images follow the single-frame pathway; SSv2 retains ordered video tokens. The category-specific and unified evaluations retain their own input geometry and readout settings.
InternVideo2 (Wang et al., 2024b); 2024-03	Component-dependent	Video foundation-model family with self-supervised video learning and subsequent multimodal alignment/instruction stages. The category-specific LP row in Table 29 is kept separate from the unified representatives; a family name alone does not identify its pretraining stage or feature interface.

B.6 Unified Representatives and VLM/MLLM References

The unified Basic Five roster contains nine configurations: CLIP, SigLIP 2 Giant, SAM 3, C-RADIOv4-H, DINOv3-7B, RAEv2-K7, V-JEPA 2.1 ViT-g, non-text VGGT-Ω, and Cosmos 3 Super. Tables 36–38 separate LP, AP, and FT. LP trains only a minimal task head; AP adds a one-block attention reader followed by the same LP task head to the frozen backbone; FT updates the backbone and task head from the pretrained weights independently for each task. Inputs follow the shared 224-pixel protocol, except COCO's 800–1,333-pixel detection geometry, rather than each checkpoint's native resolution.

The planned downstream seeds are $\{0, 1, 2\}$; tables report completed repetitions, with mean and standard deviation only for repeated runs. These are downstream repetitions, not independent pretraining runs. Unified scores use the final scheduled epoch without per-model head, layer, or learning-rate selection. Optimization can differ between LP/AP/FT, so cross-track changes do not isolate trainable scope alone. Whole-system parameter counts are not visual-component sizes.

Table 7: Unified Representatives and Additional References: models, supervision status, and feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
DINOv3-7B (Siméoni et al., 2026); 2025-08	Strict vision-native	`facebook/dinov3-vit7b16-pretrain-lvd1689m`, ViT-7B/16 with four registers. Last-layer CLS features are L2-normalized for global readouts; dense readouts exclude CLS/register tokens. Unified inputs are 224 pixels. This is neither the local ViT-B CTRL model nor a distilled student.
VGGT-Ω (Wang et al., 2026a); 2026-05	Strict vision-native	Non-text `vggt_omega_1b_512.pt`, at protocol resolution 224 rather than the native 512. Block-23 frame/global patch features are concatenated (2048-D); one camera and 16 register tokens are removed. Global tasks mean-pool patches; dense tasks keep their spatial grid. Images use $S = 1$, SSv2 16 ordered frames in one aggregator pass. Prediction heads and text-aligned weights are unused.

Continued on the next page

Table 7 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
RAEv2 $K{=}7$ tokenizer (Singh et al., 2026); 2026-05	Strict vision-native	`facebook/dinov3-vitl16-pretrain-lvd1689m` in the official `dinov3l-k7` tokenizer. The seven-layer aggregation is defined in Table 5; CLS and four registers are excluded. Global features are spatially pooled and L2-normalized; dense features retain the patch grid. Inputs are 224 pixels and SSv2 is frame-wise. No decoder, DiT, or generator latent scaling is used.
V-JEPA 2.1, ViT-g (Mur-Labadia et al., 2026); 2026-03	Strict vision-native	`vjepa2_1_vitg_384.pt`, key `target_encoder`, final `norms_block[-1]` patch features. No CLS/register tokens are present. Images use the single-frame patch embedding without frame duplication; SSv2 uses 16 frames with tubelet size 2. Unified 224-pixel inputs use rotary-position interpolation. Neither the ViT-L student nor the action-conditioned checkpoint is used.
SigLIP 2 Giant-Opt/16@384 (Tschannen et al., 2025); 2025-02	Reference	`google/siglip2-giant-opt-patch16-384`. Global features use the 1536-D attention-pooling output; dense tasks use patch features. The unified protocol uses 224-pixel inputs with positional interpolation. Text features are not used at evaluation, but pretraining remains vision–language. So400m is a separate checkpoint.
C-RADIOv4-H	Reference	`nvidia/C-RADIOv4-H`: ViT-H/16 student distilled from SigLIP 2-g, DINOv3-7B, and SAM 3. Global readouts concatenate the SigLIP/DINO summary tokens (2560-D); spatial readouts use 1280-D patches after dropping three teacher CLS and seven register tokens. Unified inputs are 224 pixels. Teachers are not loaded at evaluation, but their supervision provenance remains. The repeated category entry refers to the unified FT result rather than a new run.
Cosmos 3 Super (NVIDIA, 2026); 2026-06	Reference	`nvidia/Cosmos3-Super`: vision encoder plus official spatial merger, not the pre-merger-only path. Global readouts mean-pool final 5120-D `pooler_output` tokens; dense readouts use their official $H/32 \times W/32$ raster order. Inputs follow protocol resolution with padding to multiples of 32. Still images are duplicated for temporal packing; SSv2 encodes 16 frames this way. LP/FT pool spatially then temporally; AP reads 16 frame tokens. DeepStack, the language/world-model stacks, and the VAE are unused. The separate 448-pixel pre-merger component is 1152-D, not this interface.
CLIP ViT-L/14@336 (Radford et al., 2021); 2021-03	Reference	`openai/clip-vit-large-patch14-336`. Unified 224-pixel inputs use positional interpolation. Global readouts use projected 768-D CLS features with L2 normalization; dense readouts use 1024-D patch features. The text tower and logit scale are unused for extraction. The category-specific evaluation is separate.
SigLIP 2 So400m/14@384 (Tschannen et al., 2025); 2025-02	Reference	Additional category-specific reference: `google/siglip2-so400m-patch14-384`. Distinct from the Giant row above.
SAM 3 ViT (Carion et al., 2026); 2025-11	Reference	`facebook/sam3`, ViTDet-L/14 visual trunk only. Final 1024-D patch features are spatially pooled globally or retained densely; there is no CLS/register token. The 224-pixel protocol uses the documented positional-grid crop and recomputed axial rotary positions. Text, prompt, mask, tracker, and memory modules are not used. The category-specific evaluation remains separate.

B.7 Additional VLM/MLLM, Multimodal, and Other Models

Table 8 completes the category-specific roster in Table 35; models already cataloged in Table 7 are not repeated. Molmo-7B-D and Molmo-7B-O use visual-token states inside their language backbones, not raw CLIP features. Category placement alone does not establish language supervision: the evaluated Sapiens2 pretraining backbone is vision-native, whereas the SAM segmentation checkpoints and Qwen3-VL are references. The additional systems are not assigned the unified nine-model recipe by category membership. Their Basic Five results measure representation transfer rather than performance with their original task decoders or language-generation outputs.

Table 8: Additional VLM/MLLM, Multimodal, and Other Models: model identity and category-specific feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
SAM 1, ViT-H (Kirillov et al., 2023); 2023-04	Reference	Mask-supervised promptable segmentation model. BASIC5 LINEAR, seed 0, uses official `sam_vit_h_4b8939` and the 256-D ImageEncoderViT neck map. Global features are spatial means. The prompt encoder, mask decoder, prompts, and output masks are excluded; SSv2 averages frame-wise features temporally.
SAM 2.1, Hiera-Large (Ravi et al., 2025); 2.1 release	Reference	Image/video segmentation checkpoint `sam2.1_hiera_large.pt`. BASIC5 LINEAR, seed 0, extracts the final 1152-D Hiera trunk stage (7×7 at 224 pixels), excluding FPN neck, prompt/mask modules, and memory. SSv2 averages 16 independently encoded frame representations.
SAM 3.1, ViTDet-L/14 (Carion et al., 2026; Meta AI, 2026); 2026-03	Reference	Official Object Multiplex checkpoint `sam3.1_multiplex.pt`. BASIC5 LINEAR, seed 0, uses final 1024-D visual-trunk patches, with no CLS token; global features are spatial means. Neck/FPN, DETR, mask, multiplex, and text modules are excluded. SSv2 uses frame-wise extraction followed by temporal averaging.
SAM 3D (Meta AI, 2025); 2025-11	Component-dependent	Image-to-3D model family with separate Objects and Body releases. The category-specific result is retained under its recorded SAM 3D identity, separately from SAM 1/2.1/3/3.1. It is not treated as evidence that both 3D releases, or their full reconstruction systems, were evaluated.
FLUX2 (Black Forest Labs, 2025); 2025	Component-dependent	FLUX.2 is a language-conditioned image-generation and editing family with separate autoencoder and transformer components. The recorded FLUX2 transfer result is distinct from the DiT/RAE component experiments; no particular release, generator block, or diffusion time is inferred from the family name.
Sapiens2-1B pretrain (Khirodkar et al., 2026); 2026-04	Strict vision-native	Self-supervised human-image pretraining combining masked reconstruction and self-distillation. BASIC5 LINEAR, seed 0, uses the official 1B pretraining backbone and its last-layer 1536-D patches after `ln1`, dropping CLS and eight register tokens. Pose, segmentation, normal, pointmap, and matting heads are excluded. SSv2 uses frame-wise features followed by temporal averaging.
Qwen3-VL-8B (Bai et al., 2025)	Reference	Eight-billion-parameter dense vision–language model family for image, video, and text understanding. The category-specific Basic Five row is a representation-transfer comparison, not a zero-shot text-prompt or generated-answer benchmark. It remains separate from CLIP/SigLIP and is not assigned their visual-token interfaces or the unified LP/AP/FT settings.

Continued on the next page

Table 8 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
Molmo-7B-D (Deitke et al., 2025)	Reference	Official `allenai/Molmo-7B-D-0924`: CLIP ViT-L/14-336, connector, and Qwen2-7B. BASIC5 LINEAR, seed 0, reads 3584-D visual-token residuals at block 27, before `ln_f`; dense heads use the crop-0 12×12 grid. Raw CLIP features, text tokens, and logits are excluded. SSv2 averages frame-wise representations temporally.
Molmo-7B-O (Deitke et al., 2025)	Reference	Official `allenai/Molmo-7B-O-0924`: the same CLIP/connector with OLMo-7B. BASIC5 LINEAR, seed 0, reads 4096-D visual-token residuals at block 31 before `ln_f`, with an empty prompt. Dense heads use the crop-0 12×12 grid; SSv2 uses frame-wise extraction and temporal averaging.

B.8 Video Generative and Multimodal World Models

DINO-WM, Vid2World, and AlayaWorld v1.1 supply the world-model representation comparisons in Table 30 and have individual catalog entries below. These are downstream representation results, not rollout or action-generation evaluations. The component studies below separately identify VAE, generator-input, internal-block, and visual-encoder paths. COCO multilabel classification remains distinct from object detection. Null text at extraction does not establish strict vision-native pretraining.

Table 9: Video Generative and Multimodal World Models: model families and component-specific feature configurations.

Model, Ref., and Release	Status	Description and Checkpoint
DINO-WM (Zhou et al., 2025); 2024-11	Component-dependent	Predicts future spatial DINOv2 features with action-conditioned dynamics for planning. The frozen visual encoder and the learned dynamics predictor have different supervision provenance; the model name alone does not identify which supplies a transfer feature. The reported Basic Five row is separate from action-free DINOv2 encoder evaluations.
Vid2World (Huang et al., 2026); 2025-05	Component-dependent	Adapts pretrained video diffusion to autoregressive world modeling with frame-level action conditioning. Its Basic Five entry measures downstream transfer, not interactive rollout quality. Action-conditioned dynamics and a frozen upstream encoder are not interchangeable representation paths.
AlayaWorld v1.1 (AlayaWorld Team et al., 2026); 2026-08	Component-dependent	Interactive autoregressive video world model with camera control, text prompts, and spatial/temporal memory. The v1.1 release includes an autoregressive teacher and a distilled few-step student; neither a training stage nor a component checkpoint is inferred from the version label. Its Basic Five transfer row is distinct from LTX-2.3 component evaluations.
Cosmos 3 Nano (NVIDIA, 2026); 2026-06	Component-dependent	Additional model identity in the extended evaluation (Tables 39). It is not one of the nine unified Basic Five representatives. The Super checkpoint's component dimensions, feature adapter, and LP/AP/FT settings are not transferred to this entry.
Cosmos 3 Super components (NVIDIA, 2026); 2026-06	Reference	The component campaign separates the 448-pixel pre-merger vision encoder (1152-D), merger (5120-D), DeepStack outputs at 8/16/24 (5120-D), VAE latent, generator input (48-D), and generator block/timestep configurations. Each interface has its own result identity. The component-study readouts and scaled-depth settings are not the unified Basic Five recipe; encoder/merger COCO mAP is not bounding-box detection AP.

Continued on the next page

Table 9 (continued)

Model, Ref., and Release	Status	Description and Checkpoint
LTX-2.3 Lightricks; LTX-2.3 release	Reference	Language-conditioned video generator. Features include the VAE, generator input, and blocks 0/24/47 at $t = 0.1$ with null text. Zero-, shuffled-, and noised-latent-only variants and the block-24 random-generator control test input and generator contributions.
HunyuanVideo-1.5 (Wu et al., 2025); 2025-11	Reference	The 480p text-to-video checkpoint supplies VAE and generator blocks 0/26/53 at $t = 0.1$, with zero text during extraction. The component study also includes zero-, shuffled-, noised-latent-only, and random-generator controls.
Wan2.2-TI2V-5B (Team Wan, 2025); Wan2.2 release	Reference	VAE, generator-input, and blocks 0/7/14/22/29 from the TI2V-5B release. Generator features use null text and $t \in \{0, 0.1, 0.5, 0.9\}$. The layer/timestep grid and zero-, shuffled-, noised-latent-only, and random-generator controls remain separately identified.

B.9 Experiment-Specific Settings

Evaluation tracks and metrics. ASIS/category-specific, CTRL, unified LP/AP/FT, extended, and component evaluations retain their own checkpoint, readout, preprocessing, and metric definitions. The unified primary metrics are ImageNet-1K top-1, COCO bounding-box AP, ADE20K mIoU, NYUv2 RMSE (lower is better), and SSv2 top-1 (Appendix E.4). Secondary metrics are shown only when consistently available within a table, and missing results are not zeros.

Unified feature adaptation. Table 7 identifies global/spatial features and excluded special tokens. Unified SSv2 uses 16 frames: image encoders operate frame-wise, while V-JEPA and VGGT retain their temporal paths. Padding, positional interpolation, and temporal packing adapt each checkpoint to the common geometry. Unused text, geometric, and generative output heads do not supply transfer features. Detailed LP/AP/FT are in Appendix E.

Instance Discrimination. Tables 18 and 19 separate design ablations from capacity/data scaling. The single-view ViT-B/16 baseline uses the shared ImageNet-1K/300-epoch pretraining recipe and a 128-D normalized projection, memory-bank NCE ($\tau = 0.07$, 4,096 negatives), and bank momentum 0.5; bank entries and the NCE normalization constant are synchronized across ranks. Global crops are 224 pixels at scale 0.2–1.0, with horizontal flip, grayscale probability 0.2, and color jitter 0.4. Two-view training averages the two NCE losses and updates the bank with the normalized mean of the global-view projections. Multi-crop adds four 96-pixel crops at scale 0.05–0.32, averages all six losses, and updates the bank from the globals only. It is not the two-view-only configuration.

The remaining ablations modify that two-view baseline individually: S1 excludes the five nearest negatives from a sampled set of $4{,}096 + 5$ after epoch 20; S2 changes bank momentum to 0.99 without adding a teacher; S3 adds a 0.1-weighted, view-averaged KoLeo term over the global batch; A1 uses four prototypes per instance with maximum-similarity scoring and nearest-slot updates; A2 adds 0.1-weighted NCE on four patch tokens per view using the shared projector and image-instance bank. Downstream seed-0 probes use frozen backbone CLS/patch features, not the 128-D projection or bank; depth uses per-image median scaling. ResNet-50 ASIS is a separate historical reference. ViT-L and ImageNet-21K runs additionally change capacity, data, or duration, as specified per row.

DINOv3. The head-design and hyperparameter series (Tables 15) use frozen epoch-300 EMA teachers of ImageNet-1K-pretrained ViT-B/16 with four registers. H1 shares the MLP and prototypes between DINO CLS and iBOT patches; H1P shares only prototypes and H1M only the MLP. H1S200 splits the shared head at epoch 200; H1CORE resumes from that point without Gram anchoring. H1TA adds identity-initialized token-type affine parameters before the shared head, leaving KoLeo/Gram features unchanged. H1JA jointly balances CLS/patch teacher assignments with equal mass. Gram anchoring otherwise runs during epochs 251–300.

Local architecture/data comparisons are separate from public-checkpoint scaling (Tables 16 and 17). Public ViT-B/L/H/7B scaling reports mean and standard deviation over three downstream seeds, not three pretrained backbones. The repeated 7B LP scores appear once in Table 36.

Other classical methods and comparison scope. Tables 20–22 report Colorization, Rotation, and SplitBrain. ImageNet-1K trajectories are evaluated at epochs 100/200/300; ImageNet-21K Winter 2021 runs at 30/60/90 (AdamW 6×10^{-4}, batch 1,024, pretraining seed 42). Colorization uses Lab $L/100$ input; Rotation predicts $0/90/180/270$ degrees. SplitBrain predicts Lab $L \leftrightarrow ab$ with 50/313-bin cross-entropy losses; its two streams have widths 384 (ViT-B) or 512 (ViT-L), concatenated to 768/1024-D features. ImageNet-21K extraction normalizes L as $(L-50)/50$ and ab by 128. These comparisons change data exposure, duration, and sometimes readouts, rather than isolating data scale at matched compute.

Extended evaluation. Table 39 contains extended evaluation with six model configurations, retained as protocol-specific reference records rather than unified LP/AP/FT results. Task directions are respected; heterogeneous raw scores are not averaged into a common accuracy.

Component and input interventions. Layer, diffusion timestep, and class/text conditioning identify each generator-feature configuration. Zero-, shuffled-, and noised-latent controls and random generators remain separate; a shuffle leaving its input unchanged does not establish robustness. Layer/timestep maxima are not combined into a fictitious checkpoint.

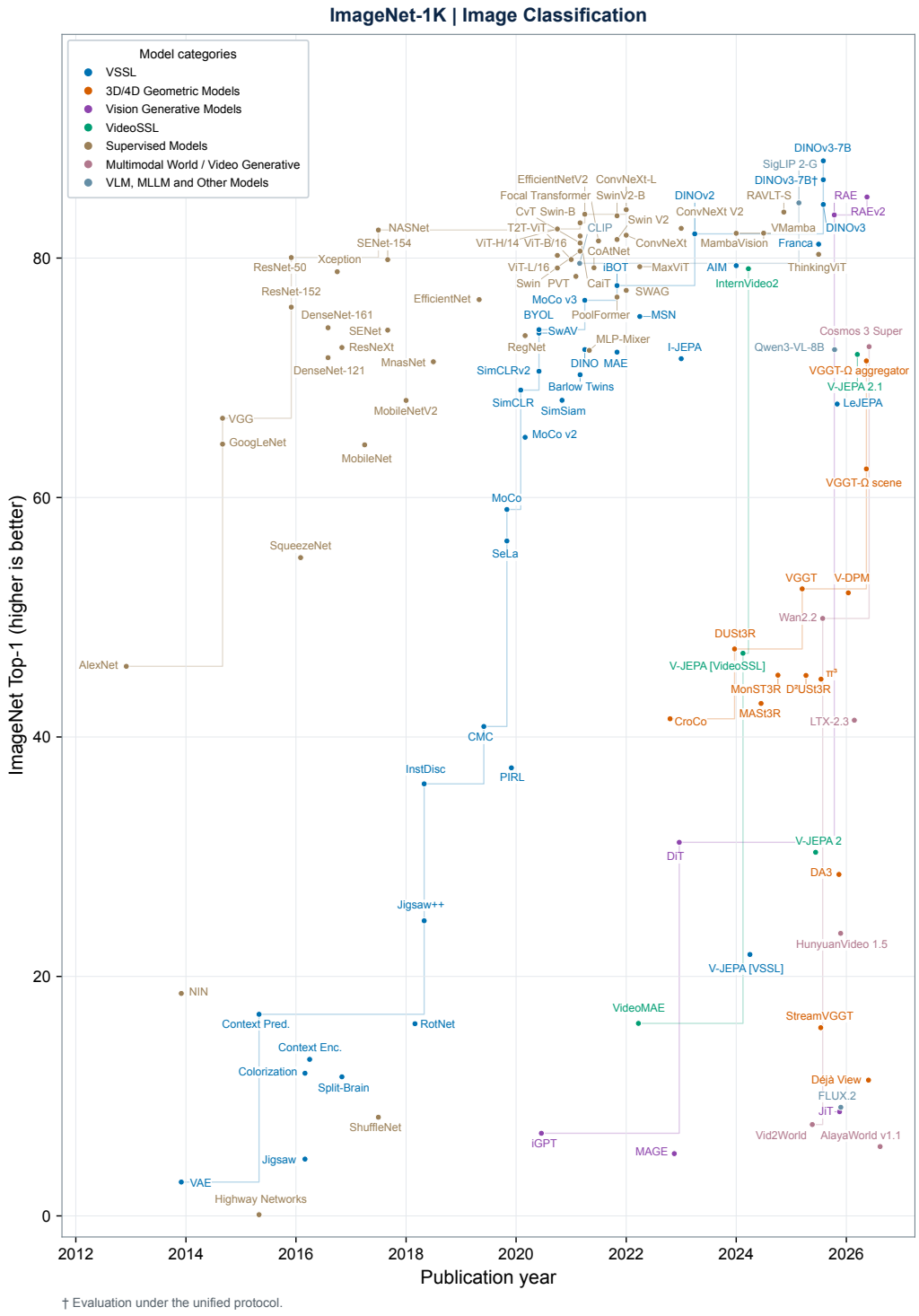
Supplementary frontier-model references. Table 42 compares GPT-5.6 Sol, GPT-6 Astra, and existing DINOv3-7B LP/AP readouts on matched 500-sample held-out subsets per dataset, without retraining DINOv3. Targets are ImageNet-1K/SSv2 top-1, COCO category-presence Micro-F1, ADE20K semantic accuracy at 32 points/image, and NYUv2 relative-depth accuracy over 20 point pairs/image (Appendix E.6). These reference-only formulations do not replace native detection, segmentation, or metric-depth evaluation and are excluded from unified rankings.

C Complete Downstream Results

We report model, interface, and ablation results with evaluation protocols kept separate. Dashes denote unavailable or source-invalidated scores, never numerical zeros. Boldface and blue shading mark the best recorded value of each metric within the stated comparison, including displayed ties; they do not indicate statistical significance. Scores use two decimals, depth errors three, and additional decimal places for small nonzero values. Reported means and standard deviations are retained where available. Within each table, companion metrics are included only when available for every row with a reported primary score. Lower NYUv2 RMSE and AbsRel are better. Repeated summaries of the same setting are listed only once; different evaluation protocols remain separate.

C.1 Historical Transfer across the Basic Five

We provide benchmark-wise, individually labeled views of the historical transfer results summarized in Figure 1. Each point represents a model at its first public release date and evaluation score. Colors identify model families, and stepwise curves show their running-best performance. Leader lines connect displaced labels to their corresponding points. Pre-training and category-specific evaluation settings differ; these plots therefore describe historical transfer rather than a unified-protocol ranking. In these figures, † denotes evaluation under the unified protocol. A reformatted Basic Five is shown below for improved readability.



† Evaluation under the unified protocol.

Figure 8: Historical transfer on ImageNet-1K for image classification. An individually labeled view of the corresponding panel in Figure 1. Performance is measured by top-1 accuracy; higher is better.

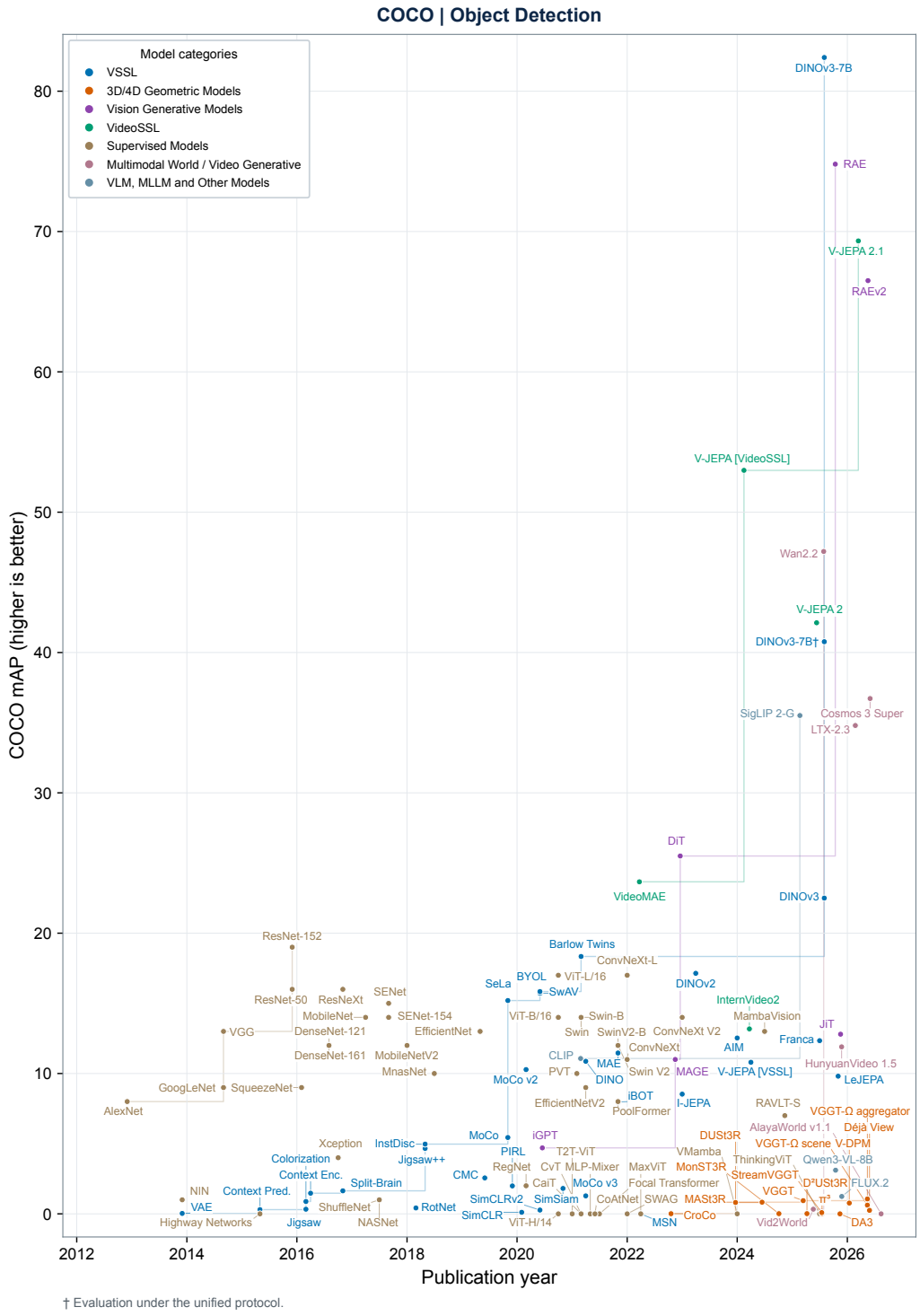


Figure 9: Historical transfer on COCO 2017 for object detection. An individually labeled view of the corresponding panel in Figure 1. Performance is measured by detection mAP; higher is better.

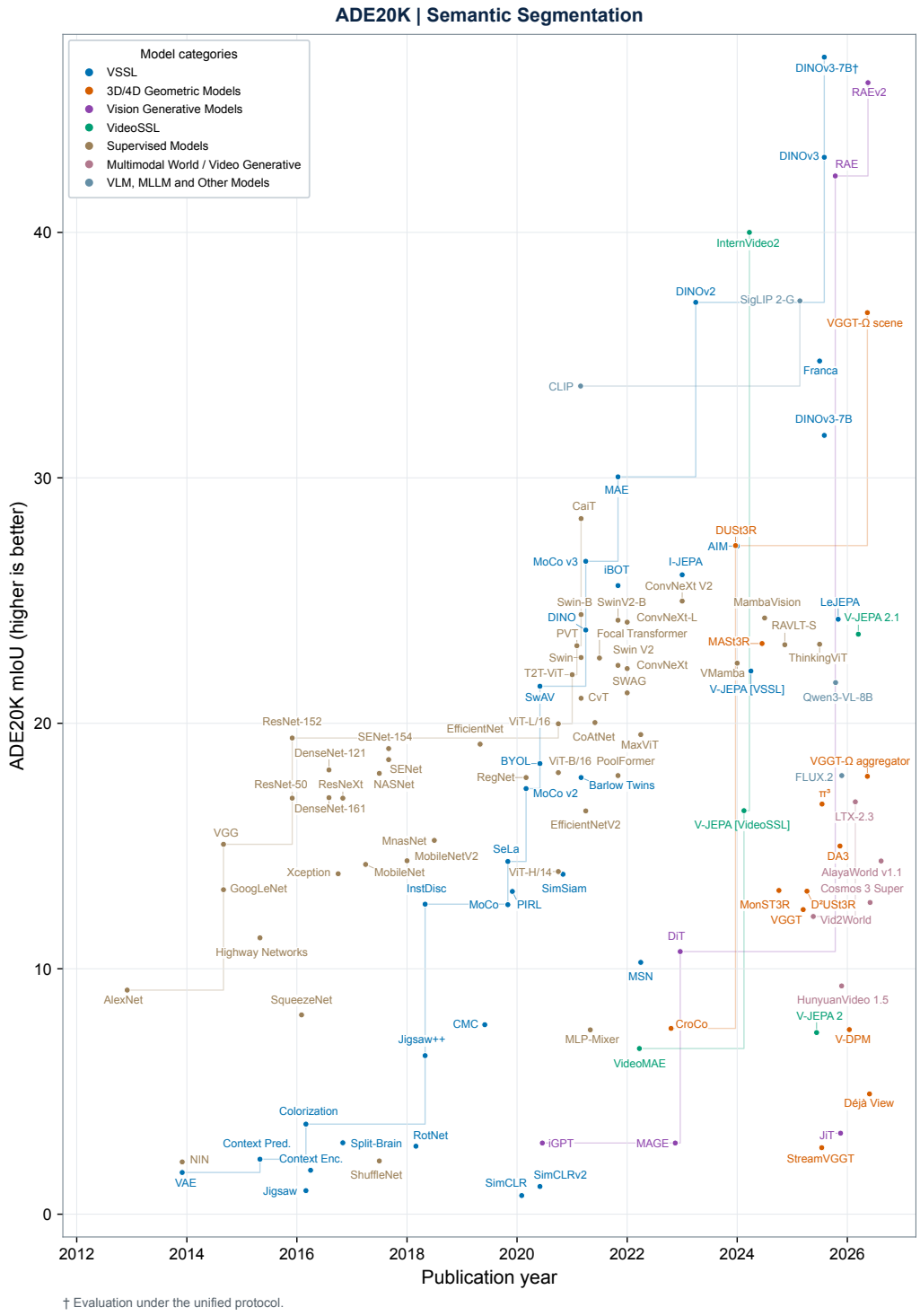


Figure 10: Historical transfer on ADE20K for semantic segmentation. An individually labeled view of the corresponding panel in Figure 1. Performance is measured by mIoU; higher is better.

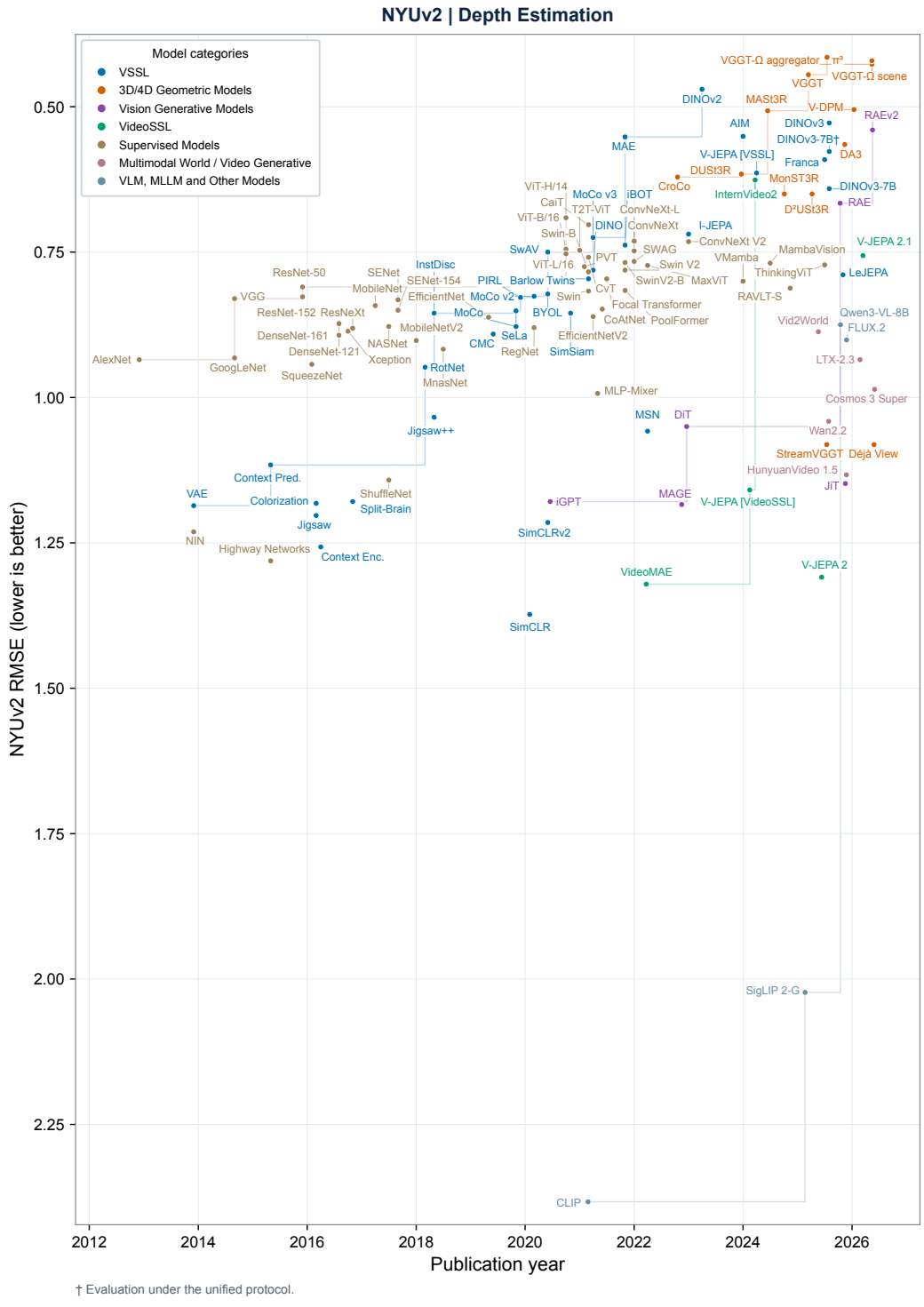
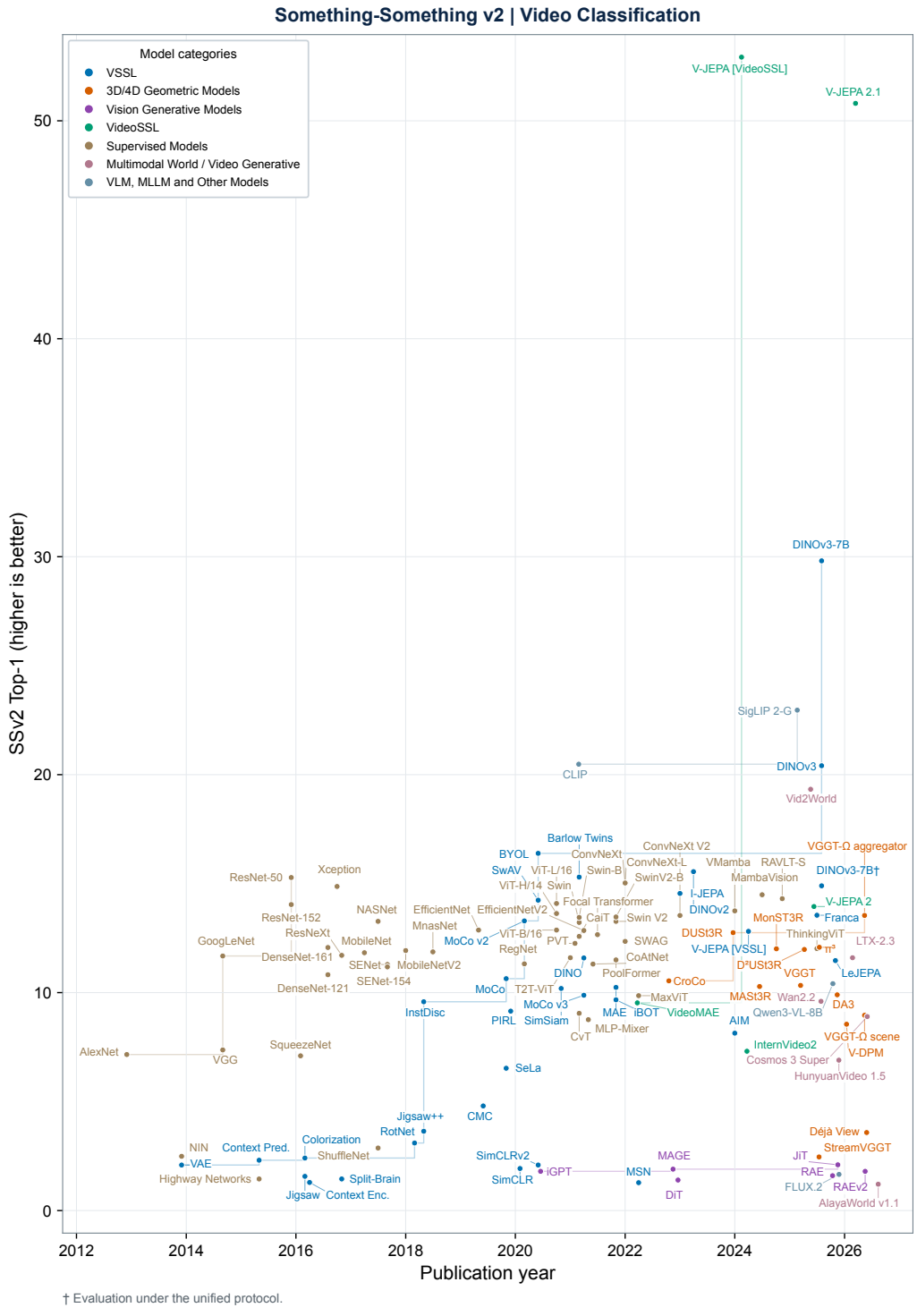


Figure 11: Historical transfer on NYUv2 for depth estimation. An individually labeled view of the corresponding panel in Figure 1. Performance is measured by RMSE; lower is better. The vertical axis is inverted so that better performance appears higher.



† Evaluation under the unified protocol.

Figure 12: Historical transfer on Something-Something V2 (SSv2) for video classification. An individually labeled view of the corresponding panel in Figure 1. Performance is measured by top-1 accuracy; higher is better.

C.2 VSSL AT ASIS

All 37 numbered methods are retained, including the five outside the controlled comparison.

Table 10: VSSL at ASIS: 37 methods on the Basic Five. Dagger-marked methods are outside the controlled comparison.

Model / setting	**ImageNet-1K**	**COCO**		**ADE20K**		**NYUv2**		**SSv2**	
	Top-1↑	mAP↑	AP50↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
VAE	2.83	0.03	0.09	1.70	35.79	1.186	0.328	2.09	8.16
Context Prediction	16.84	0.30	0.88	2.24	38.48	1.116	0.287	2.31	9.72
Colorization	11.91	0.88	2.62	3.67	44.29	1.182	0.309	2.41	9.61
Context Encoder	13.07	1.47	4.04	1.79	25.95	1.257	0.312	1.29	6.13
Jigsaw Puzzles	4.74	0.33	0.84	0.96	26.09	1.203	0.335	1.57	6.25
Rotation	16.05	0.42	1.26	2.77	40.97	0.948	0.265	3.10	11.53
DeepCluster†	5.29	—	—	—	—	—	—	—	—
SplitBrain	11.62	1.64	4.39	2.91	37.46	1.179	0.304	1.45	7.22
Jigsaw Puzzles++	24.65	4.66	11.23	6.46	48.42	1.034	0.282	3.64	13.77
Instance Discrimination	36.08	4.97	12.15	12.63	56.82	0.855	0.223	9.58	26.71
CPC†	14.84	—	—	—	—	—	—	—	—
CMC	40.88	2.56	6.90	7.72	52.14	0.891	0.239	4.80	16.99
MoCo v1	59.00	5.44	13.77	12.61	53.69	0.878	0.225	10.64	28.39
PIRL	37.42	1.99	6.10	13.15	56.10	0.828	0.216	9.15	25.71
SimCLR v1	68.97	0.11	0.34	0.76	24.32	1.373	0.326	1.93	8.27
MoCo v2	65.03	10.28	23.01	17.34	62.55	0.826	0.203	13.29	33.72
SimCLR v2	70.54	0.27	0.74	1.13	29.07	1.215	0.329	2.09	8.94
SwAV	73.73	15.72	34.03	21.51	65.36	0.750	0.189	14.24	36.30
SeLa	56.37	15.20	32.59	14.37	57.22	0.851	0.215	6.53	19.85
BYOL	74.02	15.84	33.51	18.36	61.76	0.822	0.213	16.39	39.04
SimSiam	68.12	1.81	5.30	13.85	60.65	0.855	0.206	10.19	27.42
Barlow Twins	70.25	18.34	37.06	17.79	61.74	0.796	0.208	15.30	37.68
MoCo v3	76.47	1.28	3.81	26.60	71.41	0.725	0.183	9.88	26.55
DINO	72.36	10.87	23.73	23.80	68.65	0.781	0.188	11.59	30.71
BEiT†	27.59	—	—	—	—	—	—	—	—
MAE	72.14	11.46	23.83	30.04	72.73	0.552	0.132	10.24	27.77
SimMIM†	39.46	—	—	—	—	—	—	—	—
iBOT	77.70	7.97	18.40	25.61	69.31	0.738	0.173	9.67	27.66
MSN	75.12	0.000120	0.000590	10.26	56.42	1.058	0.251	1.28	4.48
DINOv2	82.02	17.14	36.90	37.15	74.59	**0.470**	**0.109**	15.55	39.34
I-JEPA	71.59	8.53	18.76	26.05	69.12	0.719	0.172	14.55	36.84
AIM	79.37	12.53	26.48	27.22	70.39	0.551	0.127	8.14	21.00
V-JEPA	21.83	10.80	22.59	22.13	67.00	0.614	0.145	12.81	32.35
Franca	81.16	12.34	29.34	34.76	74.92	0.591	0.145	13.55	34.56
DINOv3	**84.47**	**22.50**	**44.37**	**43.06**	**79.06**	0.528	0.122	**20.41**	**46.54**
LeJEPA	67.81	9.81	21.56	24.24	69.18	0.789	0.190	11.47	31.07
NEPA†	1.57	—	—	—	—	—	—	—	—

† Excluded from the controlled comparison. AIM uses an attentive ImageNet probe in its original setting. COCO reports mAP / AP50.

C.3 VSSL AT CTRL

The 32 VSSL methods are reported at 100, 200, and 300 pre-training epochs. These are checkpoints along a training trajectory. The historical CLIP reference is reported separately.

Table 11: VSSL under CTRL at 100 pre-training epochs: 32 methods on the Basic Five.

Model / setting	**ImageNet-1K**		**COCO**		**ADE20K**		**NYUv2**		**SSv2**	
	Top-1↑	Top-5↑	mAP↑	AP50↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
VAE	5.26	13.34	1.11	2.73	5.15	47.49	1.252	0.331	3.56	13.02
Context Prediction	17.12	36.10	3.13	7.63	9.41	57.48	1.016	0.280	5.58	18.13
Colorization	11.69	27.30	1.97	5.09	9.92	58.28	1.034	0.278	6.45	19.59

Table 11: Controlled VSSL at 100 pre-training epochs: 32 methods on the Basic Five. (continued)

Model / setting	ImageNet-1K		COCO		ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	AP50↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
Context Encoder	1.04	3.53	0.000145	0.000699	0.52	23.02	1.398	0.318	1.28	4.87
Jigsaw Puzzles	4.68	12.45	0.30	0.84	3.55	45.44	1.143	0.311	3.20	11.43
Rotation	24.24	45.69	4.83	11.58	14.68	61.56	0.887	0.241	8.05	23.22
SplitBrain	17.07	35.13	3.57	9.00	16.08	63.90	1.001	0.258	6.41	19.87
Jigsaw Puzzles++	0.41	1.67	0.05	0.12	0.59	24.85	1.208	0.310	1.34	6.43
Instance Discrimination	39.93	65.55	4.17	10.36	17.35	65.63	0.993	0.242	7.84	23.23
CMC	53.25	76.11	0.00122	0.00452	17.12	63.21	0.868	0.228	7.06	20.58
MoCo v1	58.79	81.83	6.50	14.83	20.88	67.68	0.822	0.216	9.06	26.05
PIRL	56.16	80.36	6.20	14.33	21.04	68.18	0.788	0.208	9.54	26.49
SimCLR v1	53.70	78.70	1.69	4.71	9.36	57.57	1.082	0.269	4.75	15.75
MoCo v2	60.45	83.09	7.13	16.23	21.07	67.88	0.827	0.204	9.56	26.90
SimCLR v2	60.79	84.07	1.81	5.03	9.12	57.45	0.986	0.278	4.55	15.21
SwAV	59.46	83.62	0.00148	0.01	5.43	54.16	0.996	0.274	1.26	4.53
SeLa	9.75	24.08	2.14	5.64	9.47	57.91	0.988	0.264	4.84	16.46
BYOL	66.35	87.28	6.07	14.05	20.28	67.48	0.798	0.196	8.21	23.68
SimSiam	10.40	26.13	0.000781	0.00365	0.94	34.12	1.282	0.305	1.92	7.92
Barlow Twins	56.97	80.01	3.01	8.10	14.91	64.54	1.051	0.248	5.97	18.53
MoCo v3	64.75	86.03	6.09	14.05	22.42	68.34	0.803	0.222	10.16	27.82
DINO	70.24	89.38	10.45	23.09	24.26	68.73	0.745	0.188	11.95	31.34
MAE	28.11	52.02	3.64	9.12	20.89	66.93	0.793	0.205	9.26	25.64
iBOT	69.51	89.73	9.31	21.17	26.21	69.72	0.743	0.179	11.04	30.47
MSN	68.11	86.76	10.30	23.56	25.73	69.65	0.739	0.189	12.23	32.13
DINOv2	74.74	92.24	11.95	25.43	**31.67**	**73.53**	0.683	0.168	12.84	33.46
I-JEPA	56.23	79.74	5.83	13.48	19.99	66.77	0.849	0.217	11.00	30.38
AIM	35.62	60.27	4.65	11.59	16.59	62.75	0.883	0.228	7.55	22.70
V-JEPA	52.80	77.87	6.89	15.61	16.54	61.83	0.844	0.217	8.90	26.48
Franca	73.96	92.04	10.56	22.71	30.98	73.36	0.700	0.166	12.64	32.78
DINOv3	**75.19**	**92.26**	**15.76**	**30.65**	30.42	73.50	**0.681**	**0.165**	**12.94**	**33.53**
LeJEPA	64.88	86.08	0.000868	0.00372	23.59	68.56	0.830	0.197	10.75	29.73

COCO reports mAP / AP50. Nonzero near-floor scores retain additional decimal places. LeJEPA COCO uses the recorded Faster R-CNN head learning rate of 5×10^{-4}.

Table 12: VSSL under CTRL at 200 pre-training epochs: 32 methods on the Basic Five.

Model / setting	ImageNet-1K		COCO		ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	AP50↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
VAE	5.20	13.19	0.92	2.28	5.39	48.13	1.259	0.319	3.48	12.77
Context Prediction	18.27	38.28	3.19	7.83	9.99	59.21	0.948	0.248	6.34	19.72
Colorization	3.19	9.69	1.61	4.39	10.34	59.53	1.011	0.267	6.31	18.77
Context Encoder	1.12	3.73	0.10	0.22	0.89	28.26	1.231	0.327	1.33	6.47
Jigsaw Puzzles	4.23	11.94	0.51	1.50	3.18	44.26	1.233	0.303	2.95	10.63
Rotation	23.14	44.43	5.04	12.01	14.22	61.28	0.893	0.235	8.09	23.69
SplitBrain	21.79	41.91	3.81	9.29	16.93	64.33	0.971	0.256	6.45	19.78
Jigsaw Puzzles++	0.27	1.14	0.03	0.07	0.48	23.59	1.201	0.313	1.33	6.60
Instance Discrimination	47.97	73.57	4.88	11.72	20.86	68.53	0.803	0.206	8.92	24.91
CMC	60.97	82.47	0.05	0.16	20.01	65.38	0.808	0.212	7.77	23.16
MoCo v1	63.14	85.28	6.78	15.43	22.82	69.11	0.780	0.203	9.86	27.15
PIRL	60.54	83.80	6.80	15.36	22.93	69.54	0.757	0.194	10.28	27.59
SimCLR v1	55.91	80.79	0.00114	0.01	5.47	54.13	1.052	0.287	2.64	10.82
MoCo v2	65.73	86.90	6.79	15.37	22.57	69.28	0.781	0.189	10.05	27.55
SimCLR v2	64.54	86.78	0.00110	0.01	5.45	53.25	1.091	0.272	2.72	10.28
SwAV	48.22	75.25	0.000123	0.000552	0.42	9.71	1.411	0.319	0.48	4.29
SeLa	17.37	37.28	3.38	8.79	14.00	62.13	0.916	0.260	6.57	20.61
BYOL	67.16	87.42	5.90	13.69	19.31	66.85	0.835	0.196	7.35	21.29
SimSiam	3.46	10.89	0.00164	0.01	0.63	30.36	1.276	0.322	1.45	6.57
Barlow Twins	61.77	83.69	0.00175	0.01	8.29	60.78	0.998	0.249	1.98	10.71
MoCo v3	68.01	88.39	5.54	13.32	22.92	69.26	0.791	0.200	10.81	28.60
DINO	68.98	88.44	10.03	22.33	23.20	67.85	0.776	0.185	11.06	29.74

Table 12: VSSL under CTRL at 200 pre-training epochs: 32 methods on the Basic Five. (continued)

Model / setting	ImageNet-1K		COCO		ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	AP50↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
MAE	32.95	58.19	2.62	6.58	22.95	68.45	0.709	0.185	10.48	27.63
iBOT	73.53	91.98	8.62	20.81	24.93	67.51	0.741	0.179	11.16	30.31
MSN	66.79	85.11	9.84	22.58	25.64	69.58	0.779	0.187	12.18	31.88
DINOv2	72.56	90.88	10.95	23.88	31.14	73.34	0.674	0.168	12.69	32.60
I-JEPA	59.05	81.83	6.05	13.97	20.96	67.16	0.804	0.208	12.71	33.74
AIM	34.67	58.80	4.24	10.61	17.27	63.37	0.863	0.218	7.23	22.04
V-JEPA	54.90	79.22	6.03	14.28	16.39	61.47	0.886	0.213	8.04	24.65
Franca	73.34	91.53	11.79	25.24	31.25	73.21	0.663	0.165	12.89	33.08
DINOv3	**76.92**	**93.19**	**17.62**	**33.76**	**32.30**	**73.93**	**0.661**	**0.161**	**14.08**	**34.85**
LeJEPA	69.35	88.39	0.00106	0.00422	25.62	69.95	0.758	0.195	11.29	31.03

COCO reports mAP / AP50. Nonzero near-floor scores retain additional decimal places. LeJEPA COCO uses the recorded Faster R-CNN head learning rate of 5×10^{-4}.

Table 13: VSSL under CTRL at 300 pre-training epochs: 32 methods on the Basic Five.

Model / setting	ImageNet-1K		COCO		ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	AP50↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
VAE	5.00	12.76	0.23	0.57	4.82	47.54	1.192	0.304	3.08	11.68
Context Prediction	19.11	39.13	2.86	7.21	9.38	59.55	0.953	0.242	6.07	18.40
Colorization	2.30	7.38	1.39	3.92	10.18	59.79	1.004	0.271	5.82	18.17
Context Encoder	1.09	3.65	0.06	0.16	0.93	28.65	1.174	0.312	1.34	6.45
Jigsaw Puzzles	4.08	11.75	0.54	1.47	3.08	44.35	1.189	0.286	2.60	9.97
Rotation	23.02	44.40	5.48	12.88	14.83	61.71	0.890	0.240	8.52	24.46
SplitBrain	25.79	46.99	3.39	8.97	16.94	64.46	0.976	0.257	6.17	18.84
Jigsaw Puzzles++	0.19	0.76	0.03	0.08	0.50	23.68	1.220	0.306	1.35	6.57
Instance Discrimination	51.22	76.58	4.76	11.63	22.16	69.35	0.785	0.204	9.51	25.69
CMC	63.28	84.45	0.00114	0.00481	18.00	64.64	0.861	0.215	6.32	18.95
MoCo v1	64.12	86.20	6.76	15.37	23.69	69.92	0.767	0.186	9.96	27.03
PIRL	61.81	84.83	6.47	14.84	23.83	70.08	0.720	0.183	10.37	27.89
SimCLR v1	55.79	80.59	0.00149	0.01	3.43	49.01	1.091	0.312	1.84	9.07
MoCo v2	67.01	87.74	6.08	14.27	22.87	69.60	0.743	0.182	9.73	26.82
SimCLR v2	66.33	87.63	0.00129	0.01	3.68	48.52	1.090	0.295	1.95	9.05
SwAV	47.69	74.83	0.000108	0.000540	0.11	16.88	1.412	0.319	0.16	3.27
SeLa	24.33	47.40	4.13	10.53	16.53	64.26	0.907	0.237	7.79	22.70
BYOL	68.21	88.09	5.76	13.66	19.18	66.97	0.758	0.198	7.29	20.83
SimSiam	5.83	16.89	0.01	0.01	0.70	30.11	1.291	0.302	1.54	7.00
Barlow Twins	63.07	84.70	0.00198	0.01	7.91	60.50	0.988	0.249	1.90	10.31
MoCo v3	69.89	89.71	5.44	13.42	24.10	70.17	0.779	0.213	10.68	28.42
DINO	68.49	88.17	9.43	21.07	23.08	67.76	0.761	0.199	10.66	28.86
MAE	29.35	53.67	1.71	4.47	24.05	69.22	0.722	0.180	10.47	27.80
iBOT	73.65	92.04	9.13	21.54	24.77	66.77	0.728	0.181	11.01	29.64
MSN	65.31	83.96	6.71	16.41	25.71	69.54	0.756	0.182	11.36	30.92
DINOv2	70.73	89.56	9.84	21.74	31.35	73.21	0.683	0.160	12.17	31.80
I-JEPA	58.44	81.36	5.86	13.37	20.61	66.71	0.823	0.202	**14.05**	**35.62**
AIM	30.86	54.28	3.42	8.66	16.92	62.91	0.868	0.236	6.82	20.71
V-JEPA	54.10	78.56	5.02	12.33	16.37	61.61	0.867	0.221	7.85	24.11
Franca	71.37	90.41	11.34	24.69	31.11	73.03	0.646	0.162	12.35	32.49
DINOv3	**76.80**	**93.09**	**18.47**	**35.33**	**32.63**	**74.10**	**0.643**	**0.158**	13.82	34.91
LeJEPA	70.62	88.95	0.00104	0.00421	26.65	70.51	0.717	0.182	11.75	31.59

COCO reports mAP / AP50. Nonzero near-floor scores retain additional decimal places. LeJEPA COCO uses the recorded Faster R-CNN head learning rate of 5×10^{-4}.

Table 14: CLIP reference under the ASIS and CTRL settings.

Model / setting	ImageNet-1K		COCO		ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	AP50↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
Original	**71.91**	**92.55**	3.15	9.26	**26.24**	**67.62**	**0.743**	**0.174**	**12.63**	**33.05**
Controlled, 100 epochs	66.75	86.27	**5.67**	**13.64**	19.03	64.91	0.838	0.223	9.61	26.36
Controlled, 200 epochs	64.23	84.45	5.21	12.67	18.20	64.58	0.828	0.212	8.84	24.43
Controlled, 300 epochs	63.24	83.11	4.91	11.98	17.39	64.21	0.827	0.205	8.22	22.61

C.4 Revisiting VSSL

We report learning-design ablations, hyperparameter variations, and changes in model size or pre-training data. The head-design, controlled pre-training, and official-model scaling results are kept separate.

Table 15: DINOv3 head-design ablations on the Basic Five. Configuration identifiers are described in Appendix B. Shading and bold indicate the best value in each column among the displayed configurations.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	AP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
Original separate heads	74.06	91.82	22.83	32.03	73.97	2.360	0.681	**14.60**	35.78
Shared head; broader masking (H1)	74.65	92.22	23.99	**32.35**	**74.00**	2.361	0.679	14.40	36.29
Separate MLPs; shared prototypes (H1P)	73.73	91.71	**24.25**	31.59	73.84	2.546	0.761	13.74	34.43
Shared MLP; separate prototypes (H1M)	**75.00**	92.22	17.50	31.62	73.80	2.564	0.769	13.82	34.70
Shared then separate (H1S200)	74.27	91.71	23.46	31.88	73.80	2.542	0.757	14.09	35.12
Shared head; no Gram (H1CORE)	74.47	91.82	22.33	32.11	73.75	2.474	0.726	14.04	35.33
Shared head; token affine (H1TA)	73.68	91.52	23.74	32.00	73.83	**2.134**	**0.577**	14.38	**36.51**
Shared head; joint assignment (H1JA)	74.97	**92.29**	22.51	31.29	73.59	2.552	0.763	14.07	34.89

These experiments use ViT-B/16 with four register tokens and the frozen epoch-300 EMA teacher under BASIC5_FAIR_v1, with evaluation seed 0.

Table 16: Additional DINOv3 architecture and pre-training-data configurations.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	AP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
Shared head / ViT-L / IN-1K	77.94	93.94	26.73	35.16	74.95	**2.206**	**0.620**	16.01	39.08
Original / ViT-B / IN-21K	72.05	91.01	20.61	31.01	73.78	2.551	0.762	12.76	32.69
Shared head / ViT-B / IN-21K	74.37	92.59	20.68	34.27	74.85	2.517	0.749	14.44	35.98
Original / ViT-L / IN-21K	**81.63**	**96.06**	**26.88**	**38.90**	**76.38**	2.573	0.781	**16.54**	**39.86**
Shared head / ViT-L / IN-21K	81.14	95.87	25.21	38.66	76.02	2.599	0.789	15.05	37.34

The original and shared-head ViT-B/ImageNet-1K configurations appear in Table 15.

Table 17: DINOv3 model scaling on the Basic Five (three seeds).

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	AP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
DINOv3-B	82.28 ±0.02	96.20 ±0.00	33.84 ±0.09	43.57 ±0.06	79.10 ±0.03	0.593 ±0.000	0.165 ±0.000	13.71 ±0.15	33.88 ±0.21
DINOv3-L	85.65 ±0.01	97.51 ±0.01	37.04 ±0.13	46.03 ±0.04	79.61 ±0.08	0.609 ±0.001	0.167 ±0.000	14.30 ±0.21	34.47 ±0.10
DINOv3-H	86.31 ±0.00	97.63 ±0.01	36.43 ±0.02	45.55 ±0.16	78.76 ±0.02	0.700 ±0.002	0.195 ±0.001	14.22 ±0.18	33.94 ±0.15
DINOv3-7B	**86.54 ±0.01**	**97.80 ±0.00**	**40.77 ±0.04**	**47.14 ±0.07**	**79.78 ±0.05**	**0.577 ±0.001**	**0.159 ±0.000**	**14.90 ±0.23**	**34.52 ±0.21**

Each cell reports mean ± standard deviation over three seeds. Best means are highlighted among the four displayed configurations. This public-checkpoint series is separate from the local head-design experiments.

Table 18: Instance Discrimination learning-design ablations on the Basic Five. Configuration identifiers are described in Appendix B.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	AP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
Single-view ViT-B	43.92	70.23	6.79	18.27	67.76	0.878	0.239	7.27	21.86
Two-view (A)	50.16	75.28	6.73	20.52	68.84	0.860	0.230	7.69	22.61
False-negative masking (S1)	51.30	76.67	7.08	21.57	69.22	0.864	0.232	8.12	23.73
EMA bank (S2)	20.95	41.08	3.07	6.59	55.61	0.948	0.265	4.29	14.04
KoLeo (S3)	52.05	76.87	6.96	21.39	69.18	0.862	0.229	7.80	22.46
Multi-prototype (A1)	53.81	78.54	7.69	21.77	69.39	0.847	0.228	8.17	23.72
Dense Instance Discrimination (A2)	45.58	70.87	5.53	16.30	66.24	0.874	0.237	6.80	20.59
Multi-crop (S4)	**58.70**	**82.14**	**7.73**	**25.44**	**71.04**	**0.841**	**0.225**	**8.57**	**24.35**

All rows pre-train ViT-B/16 on ImageNet-1K for 300 epochs and use the same frozen-backbone evaluation with seed 0. The multi-crop configuration uses two global views and four local crops.

Table 19: Instance Discrimination: multi-crop model-size, pre-training-data, and checkpoint comparisons.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	AP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
ViT-L / IN-1K / 100 ep	50.35	74.71	**10.54**	25.47	70.75	0.839	0.221	9.95	27.53
ViT-L / IN-1K / 200 ep	62.48	84.05	10.48	28.50	**72.10**	0.818	0.211	**10.48**	**28.97**
ViT-L / IN-1K / 300 ep	**65.20**	**85.54**	9.87	**28.86**	**72.10**	**0.805**	**0.208**	10.16	28.15
ViT-B / IN-21K / 30 ep	22.65	44.80	5.19	11.89	63.59	0.918	0.254	5.46	17.22
ViT-B / IN-21K / 60 ep	39.78	66.08	7.05	21.61	69.94	0.855	0.229	7.60	22.44
ViT-L / IN-21K / 30 ep	23.40	45.55	5.57	12.64	64.63	0.898	0.246	5.78	18.25

The single-view and multi-crop ViT-B/ImageNet-1K results at 300 epochs appear in Table 18.

Table 20: Colorization: ImageNet-21K pre-training with ViT-B and ViT-L at three checkpoints, evaluated on image classification, object detection, and semantic segmentation.

Model / setting	ImageNet-1K		COCO		ADE20K	
	Top-1↑	Top-5↑	AP↑	AP50↑	mIoU↑	pAcc↑
ViT-B / IN-21K / 30 ep	5.94	15.71	4.46	11.24	12.03	61.97
ViT-B / IN-21K / 60 ep	8.03	19.91	4.21	10.75	13.00	63.87
ViT-B / IN-21K / 90 ep	7.76	19.24	2.38	6.57	10.52	63.29
ViT-L / IN-21K / 30 ep	6.38	16.52	**5.09**	**12.51**	13.36	63.35
ViT-L / IN-21K / 60 ep	10.25	24.19	4.62	11.76	**14.71**	**65.18**
ViT-L / IN-21K / 90 ep	**11.98**	**27.26**	2.66	7.40	11.71	64.50

The ImageNet-1K trajectories at 100, 200, and 300 epochs appear in Tables 11, 12, and 13.

Table 21: Rotation: ImageNet-21K pre-training with ViT-B and ViT-L at three checkpoints.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	AP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
ViT-B / IN-21K / 30 ep	30.87	55.51	8.76	12.13	60.70	0.871	0.232	8.12	22.76
ViT-B / IN-21K / 60 ep	**33.13**	**57.79**	9.28	13.18	61.21	0.848	0.222	8.90	24.25
ViT-B / IN-21K / 90 ep	31.29	55.79	9.98	14.31	61.78	0.839	0.218	**9.69**	**25.97**
ViT-L / IN-21K / 30 ep	31.76	56.10	9.29	13.38	62.03	0.843	0.220	8.85	24.17
ViT-L / IN-21K / 60 ep	30.29	54.27	9.08	13.18	61.86	0.824	0.212	8.83	23.72
ViT-L / IN-21K / 90 ep	26.78	50.13	**10.32**	**14.98**	**62.73**	**0.807**	**0.204**	8.90	22.97

The ImageNet-1K trajectories at 100, 200, and 300 epochs appear in Tables 11, 12, and 13.

Table 22: SplitBrain: ImageNet-21K pre-training with ViT-B and ViT-L at three checkpoints.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	AP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
ViT-B / IN-21K / 30 ep	12.03	26.82	6.91	15.65	64.95	0.927	0.255	3.97	14.34
ViT-B / IN-21K / 60 ep	10.35	23.55	7.95	17.86	66.23	0.926	0.253	**4.67**	**16.21**
ViT-B / IN-21K / 90 ep	4.78	12.36	3.86	17.51	66.12	0.934	0.256	1.88	7.10
ViT-L / IN-21K / 30 ep	**15.04**	**32.08**	7.71	17.89	66.92	0.910	0.248	3.80	14.65
ViT-L / IN-21K / 60 ep	8.80	21.38	**9.05**	**20.53**	**68.38**	**0.901**	**0.241**	4.44	15.39
ViT-L / IN-21K / 90 ep	2.39	7.31	2.13	19.92	67.54	0.908	0.246	1.79	4.11

The ImageNet-1K trajectories at 100, 200, and 300 epochs appear in Tables 11, 12, and 13.

C.5 Reference: Supervised Models

The supervised-model reference category follows the model and architecture configurations in the comparison. Related architecture variants are grouped together.

Table 23: Supervised models on the Basic Five.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
LeNet-5	1.13	3.80	0.01	0.38	18.70	1.263	0.393	2.03	7.91
AlexNet	45.90	68.78	7.80	9.13	50.17	0.935	0.273	7.16	20.94
NiN	18.58	36.32	1.16	2.13	33.75	1.231	0.297	2.49	10.00
VGG	66.62	86.93	12.92	15.07	56.84	0.830	0.239	7.37	20.91
GoogLeNet	64.46	85.75	9.17	13.22	54.82	0.932	0.270	11.68	30.50
HighwayNet	0.10	0.50	0.0000220	11.26	16.88	1.281	0.366	1.45	5.88
ResNet-50	80.04	94.98	15.91	16.95	57.74	0.810	0.211	15.28	37.50
ResNet-152	75.90	92.39	**19.34**	19.40	62.62	0.827	0.212	14.04	35.29
SqueezeNet	54.98	78.41	9.01	8.12	46.55	0.943	0.263	7.10	20.60
Xception	78.86	94.17	3.57	13.87	49.50	0.886	0.245	14.87	36.40
DenseNet-121	71.68	90.17	11.64	18.10	61.49	0.873	0.243	10.82	28.17
DenseNet-161	74.18	91.90	12.00	16.97	59.77	0.893	0.227	12.07	30.77
ResNeXt	72.53	89.52	15.63	16.95	59.71	0.881	0.252	11.71	30.83
MobileNet	64.40	85.69	13.79	14.25	54.41	0.842	0.242	11.83	31.25
MobileNetV2	68.11	88.26	12.44	14.40	57.04	0.902	0.235	11.93	31.23
ShuffleNet	8.24	19.38	1.10	2.17	35.84	1.142	0.320	2.87	10.98
NASNet	82.34	95.84	0.85	17.96	59.63	0.878	0.224	13.27	33.88
SENet	73.98	90.90	14.78	18.52	61.09	0.832	0.218	11.29	28.57
SENet-154	79.87	93.98	14.20	18.97	59.01	0.850	0.223	11.18	28.70
MnasNet	71.34	90.48	9.79	15.23	57.63	0.917	0.252	11.87	31.06
EfficientNet	76.53	93.16	13.38	19.15	61.30	0.862	0.231	12.87	32.18
EfficientNetV2	83.66	96.57	8.44	16.43	54.92	0.861	0.235	12.85	32.98
RegNet	73.52	91.60	1.66	17.79	60.16	0.880	0.236	11.32	29.18
SEER RegNetY	65.78	87.94	0.00	25.22	**69.47**	**0.619**	**0.153**	**18.42**	**43.14**
ViT-B/16	80.22	94.94	14.44	17.99	66.83	0.745	0.189	12.87	33.31
ViT-L/16	79.17	94.22	17.31	19.98	66.50	0.753	0.193	14.09	35.23
ViT-H/14	82.43	96.24	0.00	13.96	54.33	0.691	0.178	13.63	33.81
T2T-ViT	79.88	94.52	0.00	21.98	66.45	0.747	0.190	11.60	30.04
CaiT	81.84	95.79	0.00	**28.34**	68.26	0.703	0.186	12.58	32.13
PVT	78.47	94.41	9.62	23.16	65.02	0.775	0.207	12.26	31.57
CvT	81.26	95.41	0.00	21.02	62.67	0.759	0.204	9.05	22.55
Swin	80.58	95.40	13.74	22.68	63.73	0.817	0.219	13.22	33.58
Swin-B	82.94	96.33	13.67	24.43	63.99	0.784	0.214	13.45	33.31
Swin V2	81.55	95.69	11.94	22.36	62.43	0.781	0.217	13.27	33.46
Swin V2-B	83.53	96.60	12.03	24.20	64.27	0.768	0.210	13.47	33.70
Focal Transformer	81.43	95.52	0.00	22.66	62.73	0.796	0.211	12.66	32.05
MLP-Mixer	72.29	89.58	0.00	7.51	39.02	0.993	0.276	8.76	24.10
PoolFormer	76.74	93.30	7.74	17.87	60.07	0.816	0.220	11.50	30.38
CoAtNet	79.19	94.36	0.00	20.03	60.06	0.848	0.227	11.31	29.62
ConvNeXt	81.91	95.83	11.29	22.23	63.17	0.748	0.206	15.03	36.58
ConvNeXt-L	**84.04**	**96.82**	16.70	24.12	65.48	0.731	0.190	15.03	35.58
ConvNeXt V2	82.48	96.23	14.10	24.98	65.26	0.732	0.194	13.54	33.82
MaxViT	79.28	91.85	0.00	19.54	58.84	0.773	0.205	9.86	24.13
SWAG	77.30	93.58	0.00	21.24	59.21	0.766	0.204	12.35	30.90
VMamba	82.07	96.00	0.00	22.45	62.74	0.800	0.219	13.75	34.10

Table 23: Supervised models on the Basic Five. (continued)

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
MambaVision	82.08	96.08	13.19	24.29	64.82	0.769	0.200	14.49	35.13
RA-VLT	83.84	96.68	6.96	23.20	62.17	0.812	0.221	14.31	34.46
ThinkingViT	80.31	95.08	0.00	23.22	66.83	0.772	0.189	12.02	31.85

C.6 3D Geometric Models

The representative model comparison is followed immediately by VGGT-Ω feature-interface ablations.

Table 24: 3D geometric models: representative model and feature configurations.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
CroCo v2	41.52	65.00	0.01	7.57	63.26	0.621	0.152	10.54	28.77
DUSt3R	47.35	70.69	0.81	**27.24**	**73.20**	0.616	0.138	12.75	33.00
MASt3R	42.80	66.84	0.83	23.25	72.42	0.507	0.116	10.28	27.96
DA3	28.52	51.26	0.00	15.00	67.84	0.565	0.137	9.90	28.17
π^3	44.83	68.98	0.02	16.71	71.22	**0.415**	**0.085**	12.08	31.25
VGGT	52.37	75.25	0.94	12.41	68.05	0.445	0.099	10.33	28.28
VGGT-Ω / aggregator	**71.41**	**89.43**	**1.06**	17.84	71.87	0.427	0.094	**13.54**	**34.88**

Table 25: VGGT-Ω: additional feature-interface ablations.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
Text-aligned aggregator†	**72.67**	**90.49**	0.01	18.23	70.61	0.431	0.095	**13.69**	**35.46**
Scene tokens	62.39	83.12	0.62	**36.73**	76.87	0.421	0.094	8.96	25.75
Camera tokens	8.88	21.42	0.75	36.69	**76.88**	**0.420**	**0.092**	4.60	14.50
Dense tokens	10.08	23.48	**0.96**	4.11	51.44	0.834	0.197	4.60	13.92

The non-text aggregator baseline appears in Table 24. † The text-aligned aggregator is a separate reference. Best values refer only to the interfaces displayed here.

C.7 4D Geometric Models

Five 4D geometric models are evaluated on the Basic Five.

Table 26: 4D geometric models on the Basic Five.

Model / setting	ImageNet-1K	COCO	ADE20K	NYUv2		SSv2
	Top-1↑	mAP↑	mIoU↑	RMSE↓	AbsRel↓	Top-1↑
Deja View	11.34	0.25	4.90	1.081	0.322	3.58
StreamVGGT	15.72	0.09	2.71	1.081	0.317	2.46
MonST3R	45.16	0.01	**13.19**	0.650	0.157	**12.01**
D^2USt3R	45.14	0.02	13.16	0.650	0.157	11.98
V-DPM	**52.04**	**0.77**	7.52	**0.505**	**0.114**	8.55

C.8 Vision Generative Models

The representative-model comparison is followed by component, timestep, and ablation results. Identical settings already in the overview are not repeated.

Table 27: Vision generative models: selected model and interface configurations.

Model / setting	ImageNet-1K		COCO	ADE20K		NYUv2		SSv2	
	Top-1↑	Top-5↑	mAP↑	mIoU↑	pAcc↑	RMSE↓	AbsRel↓	Top-1↑	Top-5↑
iGPT-S / 32px	6.90	17.90	4.70	2.90	37.00	1.179	0.354	1.80	8.00
data2vec 2.0	41.30	66.60	19.70	14.40	69.00	0.852	0.241	3.30	12.40
MAGE / ViT-B/16	5.20	13.40	11.00	2.90	40.30	1.184	0.397	1.90	7.20
DiT-XL/2	31.20	53.50	25.50	10.70	57.50	1.050	0.344	1.40	5.50
JiT-L	8.70	20.30	12.80	3.30	44.10	1.148	0.378	2.10	8.20
AIMv2-L	82.10	96.40	33.60	17.10	70.00	0.877	0.244	**8.00**	**20.60**
RAEv1 / DINOv2-B encoder	83.60	96.40	**74.80**	42.30	77.50	0.666	0.187	1.60	6.50
RAEv2 / K7 + DiT, layer 29	**85.10**	**97.10**	66.50	**46.10**	**80.00**	**0.540**	**0.141**	1.80	5.50

RAEv1 and RAEv2 distinguish encoder/tokenizer features from generator features.

Table 28: RAEv1 and RAEv2: additional components, timesteps, and ablations. Best values are highlighted separately within each model.

Setting / component	ImageNet-1K Top-1 / Top-5	COCO mAP	ADE20K mIoU / pAcc	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
RAEv1 components and timesteps					
Tokenizer; omit CLS/registers	82.40 / 96.40	75.90	42.30 / 77.50	0.067 / 0.187	1.50 / 6.10
Generator input; latent + BN statistics	82.40 / 96.40	76.00	42.30 / 77.50	0.067 / 0.187	1.50 / **6.80**
Block 0 conditioned; t=0	82.20 / 96.30	76.40	42.30 / 77.80	0.064 / 0.177	1.40 / 5.80
Block 0 conditioned; t=0.1	82.30 / 96.30	77.30	42.30 / 77.60	0.065 / 0.181	**1.90** / 5.70
Block 0 conditioned; t=0.5	82.00 / 96.30	77.10	41.10 / 76.50	0.070 / 0.198	1.40 / 6.10
Block 0 conditioned; t=0.9	81.90 / 96.20	69.80	33.10 / 70.80	0.086 / 0.261	1.40 / 6.20
Block 6 early; t=0	82.00 / 96.20	74.70	42.80 / 78.20	0.061 / 0.168	1.40 / 5.40
Block 6 early; t=0.1	82.40 / 96.40	76.50	42.80 / 78.30	0.060 / 0.159	1.50 / 5.40
Block 6 early; t=0.5	82.20 / 96.40	**77.60**	42.40 / 77.90	0.060 / 0.162	1.30 / 5.60
Block 6 early; t=0.9	81.30 / 95.90	75.40	37.30 / 74.70	0.067 / 0.186	1.30 / 5.90
Block 13 middle; t=0	81.90 / 96.10	75.30	42.80 / 78.30	0.061 / 0.165	1.30 / 6.60
Block 13 middle; t=0.1	81.70 / 95.90	75.00	42.70 / 78.20	0.060 / 0.159	1.60 / 6.60
Block 13 middle; t=0.5	81.70 / 95.90	74.90	42.70 / 78.20	**0.059** / 0.159	1.60 / 6.60
Block 13 middle; t=0.9	80.50 / 95.50	73.90	38.60 / 75.90	0.062 / 0.166	1.30 / 5.40
Block 20 late; t=0	79.70 / 95.00	72.00	42.90 / **78.50**	**0.059** / 0.160	1.40 / 5.80
Block 20 late; t=0.1	79.50 / 94.80	67.70	42.00 / 78.30	0.060 / 0.161	1.50 / 6.00
Block 20 late; t=0.5	78.90 / 94.30	66.30	41.40 / 78.20	**0.059** / 0.158	1.50 / 5.90
Block 20 late; t=0.9	78.30 / 94.10	68.70	39.40 / 76.70	0.060 / 0.158	1.30 / 6.20
Block 27 fusion; t=0	78.50 / 94.20	51.50	41.90 / 78.10	0.060 / 0.159	1.30 / 5.90
Block 27 fusion; t=0.1	78.20 / 94.00	48.50	41.50 / 78.10	0.060 / 0.161	1.40 / 5.70
Block 27 fusion; t=0.5	78.20 / 94.10	50.30	41.80 / 78.20	**0.059** / 0.158	1.40 / 6.40
Block 27 fusion; t=0.9	78.20 / 94.10	52.20	40.50 / 77.30	**0.059** / **0.155**	1.50 / 6.20
Block 29 pre-output; t=0	62.50 / 83.60	56.60	0.10 / 16.90	0.070 / 0.179	1.10 / 5.60
Block 29 pre-output; t=0.1	**82.90** / **96.60**	70.50	43.20 / 78.20	0.062 / 0.171	1.30 / 6.00
Block 29 pre-output; t=0.5	**82.90** / **96.60**	69.20	**43.30** / 78.30	0.061 / 0.167	1.50 / 4.90
Block 29 pre-output; t=0.9	81.90 / 96.20	65.60	41.70 / 77.60	**0.059** / 0.157	1.20 / 4.80
ViT-XL decoder; layer 13	79.30 / 94.80	68.20	41.20 / 77.70	0.067 / 0.191	1.40 / 6.40
ViT-XL decoder; pre-output	71.70 / 90.10	15.50	27.90 / 69.80	0.094 / 0.294	1.40 / 6.00
RAEv1 latent and generator ablations					
Block 0; t=0.1; noised latent only	82.30 / 96.20	76.60	42.40 / 77.60	0.065 / 0.181	1.40 / 5.90
Block 0; t=0.1; zero latent	0.20 / 0.60	4.10	0.50 / 22.20	0.126 / 0.412	1.30 / 6.60
Block 0; t=0.1; shuffled latent	0.20 / 0.60	4.50	0.50 / 21.90	0.126 / 0.413	1.50 / 6.00
Block 13; t=0.1; untrained generator	81.00 / 95.70	75.60	42.10 / 77.30	0.067 / 0.189	1.50 / **6.80**
RAEv2 components and timesteps					
DINOv3-L; last-layer encoder only	**85.50** / **97.30**	75.60	**48.10** / **80.70**	0.056 / 0.149	1.50 / 6.80
DINOv3-L; official K=7 tokenizer	85.00 / 97.20	76.70	45.80 / 79.90	0.056 / 0.150	1.50 / 6.80
Generator input; K=7 latent + statistics	85.10 / 97.20	69.10	45.70 / 79.70	0.058 / 0.157	1.50 / **6.90**
Block 0 conditioned; t=0	84.80 / 97.10	78.10	45.50 / 79.60	0.057 / 0.155	1.50 / 6.10
Block 0 conditioned; t=0.1	84.90 / 97.10	78.30	45.50 / 79.60	0.058 / 0.156	1.50 / 6.40
Block 0 conditioned; t=0.5	85.00 / 97.00	77.70	44.40 / 78.90	0.062 / 0.172	1.40 / 6.30
Block 0 conditioned; t=0.9	84.40 / 97.00	70.30	38.10 / 75.40	0.078 / 0.231	**1.60** / 6.60
Block 13 middle; t=0	84.50 / 97.00	72.50	46.00 / 80.10	0.055 / 0.144	1.30 / 5.80
Block 13 middle; t=0.1	84.50 / 96.90	71.80	46.00 / 80.10	0.055 / 0.143	1.40 / 5.90
Block 13 middle; t=0.5	84.20 / 96.90	71.30	45.80 / 80.10	0.054 / 0.138	1.30 / 6.40
Block 13 middle; t=0.9	83.70 / 96.60	74.60	39.80 / 77.30	0.057 / 0.150	1.10 / 6.30
Block 20 late; t=0	83.90 / 96.70	57.30	45.00 / 79.80	0.056 / 0.145	—
Block 20 late; t=0.1	83.80 / 96.70	56.20	44.90 / 79.80	0.055 / 0.144	1.40 / 5.90
Block 20 late; t=0.5	83.50 / 96.50	60.50	44.40 / 79.60	0.055 / 0.143	—
Block 20 late; t=0.9	83.20 / 96.30	66.90	39.90 / 77.30	0.057 / 0.152	1.30 / 6.20
Block 29 pre-output; t=0	85.10 / 97.20	67.10	46.10 / 79.90	0.054 / 0.143	—
Block 29 pre-output; t=0.5	84.90 / 97.20	64.90	46.30 / 80.30	**0.052** / **0.136**	—
Block 29 pre-output; t=0.9	83.90 / 96.70	60.30	41.80 / 78.10	0.053 / 0.138	1.20 / 4.00

Table 28: RAEv1 and RAEv2: additional components, timesteps, and ablations. Best values are highlighted separately within each model. (continued)

Setting / component	ImageNet-1K Top-1 / Top-5	COCO mAP	ADE20K mIoU / pAcc	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
ViT-XL decoder; layer 13	68.10 / 86.50	44.20	20.00 / 65.10	0.100 / 0.315	1.40 / 5.90
ViT-XL decoder; pre-output	57.70 / 78.60	10.50	11.90 / 59.50	0.110 / 0.355	1.40 / 6.10
RAEv2 latent and generator ablations					
Block 0; t=0.1; noised latent only	84.70 / 97.10	**78.80**	45.50 / 79.60	0.058 / 0.156	1.40 / 6.30
Block 0; t=0.1; zero latent	0.30 / 0.60	4.20	0.50 / 22.20	0.126 / 0.412	1.30 / 6.00
Block 0; t=0.1; shuffled latent	0.20 / 0.60	4.80	0.40 / 20.20	0.127 / 0.415	1.40 / 6.30
Block 13; t=0.1; untrained generator	83.60 / 96.40	73.90	43.40 / 78.50	0.065 / 0.184	**1.60** / 6.80

The selected RAEv1 encoder and RAEv2 layer-29 setting at $t = 0.1$ appear in Table 27. Depth values are reported as recorded for each component series; no cross-series ranking is implied.

C.9 VIDEO SSL

Table 29: Video self-supervised models.

Model / setting	ImageNet-1K Top-1↑	COCO mAP↑	ADE20K mIoU↑	NYUv2 RMSE↓	SSv2 Top-1↑
Shuffle and Learn / ResNet-50	8.12	—	—	—	—
VideoMAE / ViT-B	16.08	23.66	6.75	1.321	9.53
VideoMAE V2	5.80	6.53	12.36	0.913	2.61
InternVideo2	**79.11**	13.18	**40.00**	**0.626**	7.31
V-JEPA 1 / ViT-L	46.99	52.98	16.44	1.159	52.92
V-JEPA 2 / ViT-H	30.37	42.12	7.40	1.309	13.95
V-JEPA 2-AC / ViT-g/16	73.87	61.16	11.25	0.976	**53.54**
V-JEPA 2.1 / ViT-L	72.64	65.95	23.19	0.786	49.85
V-JEPA 2.1 / ViT-g	71.95	**69.33**	23.63	0.756	50.80

V-JEPA 2-AC is an action-conditioned reference; InternVideo2 remains a category-specific configuration rather than a unified representative.

C.10 REFERENCE: MULTIMODAL WORLD AND VIDEO GENERATIVE MODELS

Multimodal world model references and video generative models are followed by their component and ablation tables. Task formulations and depth scales are retained within each recorded protocol.

Table 30: World-model representations and selected video-generative and multimodal interfaces: protocol-specific transfer results.

Model / setting	ImageNet-1K Top-1↑	COCO mAP↑	ADE20K mIoU↑	NYUv2 RMSE↓	SSv2 Top-1↑
World-model representations					
DINO-WM	**15.33**	**5.21**	**16.19**	1.024	2.33
Vid2World	7.63	0.33	12.13	**0.887**	**19.33**
AlayaWorld v1.1	5.79	0.00100	14.39	6.135	1.21
Video-generative and multimodal interfaces					
Wan2.2-TI2V-5B / block 14	49.90	47.20	—	0.104	9.60
LTX-2.3 / block 24	41.40	34.80	**16.80**	**0.094**	**11.60**
HunyuanVideo-1.5 / block 26	23.60	11.90	9.30	0.113	6.90
Cosmos 3 Super / vision merger	**72.60**	**70.50**	12.70	0.099	8.90

Dashes indicate unreported results. Shading and bold mark the best reported values within each group; lower is better for RMSE. Results follow their respective protocols and are not ranked across groups or jointly with the unified Basic Five evaluation. Additional component settings appear in the following tables.

Table 31: Cosmos 3 Super: additional visual interfaces, generator layers, timesteps, and ablations.

Setting / component	ImageNet-1K Top-1	COCO mAP	ADE20K mIoU / pAcc	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
Visual interfaces and generator input					
vision encoder (1152-d)	**73.10**	**67.80**	2.40† / 28.50	**0.085 / 0.260**	6.10 / 17.80
deepstack 8 (5120-d)	—	40.60	6.50 / 49.90	0.111 / 0.366	4.90 / 15.50
deepstack 16 (5120-d)	—	61.80	13.50 / **58.00**	0.095 / 0.291	**9.00 / 23.40**
deepstack 24 (5120-d)	71.90	65.60	**14.10** / 57.20	0.095 / 0.296	8.20 / 23.30
vae latent (48-d)	6.40	8.20	2.10 / 36.40	0.128 / 0.436	1.50 / 7.00
generator input (48-d)	6.30	8.20	2.10 / 36.40	0.128 / 0.437	1.50 / 7.20
Generator layers at t=0.1					
Generator layer 0; t=0.1 (5120-d)	8.80	14.00	2.00 / 28.90	0.133 / 0.491	2.00 / 8.20
Generator layer 4; t=0.1	9.70	15.00	2.00 / 30.70	0.155 / 0.512	2.00 / 8.20
Generator layer 8; t=0.1	8.90	15.10	2.40 / 35.20	0.128 / 0.400	2.00 / 8.10
Generator layer 12; t=0.1	10.10	15.40	3.50 / 38.80	0.126 / 0.410	2.10 / 8.30
Generator layer 16; t=0.1	11.80	15.70	4.00 / 40.20	0.125 / 0.421	2.10 / 8.40
Generator layer 20; t=0.1	14.80	16.70	5.20 / 44.30	0.117 / 0.396	2.20 / 8.60
Generator layer 24; t=0.1	18.70	19.30	7.10 / 47.80	0.118 / 0.413	2.50 / 9.30
Generator layer 28; t=0.1	22.50	23.20	9.10 / 51.70	0.118 / 0.407	2.90 / 10.50
Generator layer 32; t=0.1	22.20	22.80	9.20 / 51.10	0.116 / 0.379	2.90 / 10.30
Generator layer 36; t=0.1	29.30	24.90	11.30 / 54.60	0.109 / 0.359	3.60 / 12.80
Generator layer 40; t=0.1	28.80	23.50	11.00 / 53.90	0.109 / 0.338	3.90 / 13.40
Generator layer 44; t=0.1	28.90	22.40	10.60 / 54.50	0.110 / 0.365	3.80 / 12.80
Generator layer 48; t=0.1	30.40	22.70	9.60 / 53.60	0.110 / 0.367	3.50 / 12.20
Generator layer 52; t=0.1	27.50	23.40	8.40 / 50.20	0.114 / 0.371	3.50 / 11.90
Generator layer 56; t=0.1	23.20	23.60	7.50 / 50.80	0.113 / 0.377	3.50 / 12.50
Generator layer 60; t=0.1	22.90	24.10	7.50 / 50.80	0.114 / 0.381	3.60 / 12.60
Generator layer 63; t=0.1	22.70	24.30	7.70 / 51.00	0.114 / 0.393	—
Generator timesteps and controls					
Generator layer 16; t=0.0	11.60	15.80	3.90 / 40.20	0.122 / 0.410	2.10 / 8.50
Generator layer 16; t=0.5	9.70	14.50	3.10 / 39.00	0.120 / 0.389	2.10 / 8.40
Generator layer 16; t=0.9	7.30	11.70	1.60 / 28.70	0.121 / 0.404	1.70 / 7.60
Generator layer 32; t=0.0	22.70	23.00	9.20 / 51.90	0.111 / 0.368	2.90 / 10.60
Generator layer 32; t=0.5	13.90	17.80	6.40 / 47.10	0.116 / 0.395	2.80 / 10.60
Generator layer 32; t=0.9	7.50	12.40	3.00 / 40.60	0.122 / 0.410	2.20 / 8.60
Generator layer 48; t=0.0	29.80	22.80	9.70 / 53.70	0.111 / 0.371	3.30 / 11.70
Generator layer 48; t=0.5	18.20	18.30	7.10 / 50.30	0.115 / 0.394	3.80 / 12.60
Generator layer 48; t=0.9	8.40	12.40	3.10 / 40.30	0.121 / 0.408	2.50 / 9.60
shuffled latent (L32 t=0.1)	13.80	20.00	3.20 / 40.80	0.121 / 0.402	2.40 / 9.40
untrained generator (L32 t=0.1)	7.20	11.90	2.50 / 38.00	0.124 / 0.431	1.90 / 8.00

The vision-merger reference appears in Table 30. COCO encoder/merger scores are mAP; remaining components retain their recorded settings. † The encoder ADE20K entry is flagged for a token-grid mismatch and excluded from best-value highlighting.

Table 32: LTX-2.3: additional components and generator ablations.

Setting / component	ImageNet-1K Top-1 / Top-5	COCO mAP	ADE20K mIoU / pAcc	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
VAE encoder; no DiT	4.20 / 12.20	9.50	2.50 / 38.90	0.124 / 0.419	2.10 / 7.90
generator_input; no DiT	4.20 / 12.10	9.60	2.50 / 39.00	0.124 / 0.417	2.10 / 7.90
conditioned; Block 0; t=0.1; null text	12.00 / 26.10	14.60	4.80 / 47.40	0.118 / 0.382	2.50 / 9.70
pre-output; Block 47; t=0.1; null text	**30.40 / 53.50**	**25.60**	**12.20 / 59.80**	**0.102 / 0.323**	**7.20 / 20.00**
Block 0; t=0.1; zero latent	0.10 / 0.50	3.60	0.50 / 22.20	0.125 / 0.410	1.40 / 6.60
Block 0; t=0.1; shuffled latent	12.00 / 26.10	13.90	4.90 / 47.40	0.118 / 0.389	2.60 / 9.90
Block 0; t=0.1; noised latent only	12.00 / 26.10	14.60	4.90 / 47.50	0.119 / 0.387	2.50 / 9.90
Block 24; t=0.1; untrained generator	2.20 / 7.10	3.90	0.90 / 26.50	0.134 / 0.452	0.60 / 2.30

The block-24 baseline appears in Table 30. COCO uses mAP. Generator extraction uses null text. Shuffled latents are a batch-size-one no-op for ImageNet, COCO, and SSv2.

Table 33: HunyuanVideo-1.5: additional components and ablations.

Setting / component	ImageNet-1K Top-1 / Top-5	COCO mAP	ADE20K mIoU / pAcc	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
VAE encoder; no DiT	4.00 / 11.70	8.70	1.90 / 36.40	0.132 / 0.449	2.10 / 8.30
Block 0; conditioned; t=0.1	10.20 / 23.40	9.30	4.10 / 47.00	0.124 / 0.420	2.60 / 9.20
Block 53; pre-output; t=0.1	**19.30 / 37.40**	**15.30**	**7.70 / 51.50**	**0.115 / 0.381**	**4.90 / 14.80**
Block 0; t=0.1; noised latent only	10.20 / 23.30	9.10	4.10 / 47.00	0.124 / 0.420	2.50 / 9.30
Block 0; t=0.1; zero latent	0.10 / 0.60	3.80	0.50 / 22.20	0.125 / 0.411	1.20 / 6.60
Block 0; t=0.1; shuffled latent	7.90 / 18.70	5.50	4.10 / 47.00	0.124 / 0.419	2.50 / 9.20
Block 26; t=0.1; untrained generator	4.40 / 12.50	8.60	1.90 / 36.50	0.132 / 0.448	2.10 / 8.50

The block-26 baseline appears in Table 30. Generator extraction uses unconditional zero text.

Table 34: Wan2.2-TI2V-5B: additional components, timesteps, and ablations.

Setting / component	ImageNet-1K Top-1 / Top-5	COCO mAP	ADE20K mIoU	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
Components and timesteps					
vae_encoder; no DiT	6.40 / 16.50	12.40	—	0.128 / 0.428	2.10 / 8.30
generator_input; no DiT	6.40 / 16.40	12.30	—	0.130 / 0.452	2.10 / 8.30
Block 0 conditioned; t=0.0	22.10 / 41.70	21.30	—	0.113 / 0.337	5.20 / 15.30
Block 0 conditioned; t=0.1; y=null	22.10 / 41.30	20.70	—	0.113 / 0.343	5.30 / 15.40
Block 0 conditioned; t=0.5	19.20 / 37.70	19.80	—	0.112 / 0.368	5.90 / 16.60
Block 0 conditioned; t=0.9	12.10 / 27.00	14.40	—	0.123 / 0.459	5.60 / 15.90
Block 7 early; t=0.0	41.20 / 64.90	39.50	—	0.115 / 0.419	8.30 / 22.50
Block 7 early; t=0.1; y=null	42.80 / 67.00	40.50	—	0.105 / 0.346	8.80 / 23.90
Block 7 early; t=0.5	32.40 / 55.50	34.10	—	0.105 / **0.318**	**9.80 / 25.60**
Block 7 early; t=0.9	12.00 / 26.70	16.80	—	0.120 / 0.326	6.30 / 17.60
Block 14 middle; t=0.0	**46.90 / 70.80**	**44.00**	—	**0.101** / 0.343	8.90 / 24.30
Block 14 middle; t=0.5	31.30 / 54.20	33.80	—	0.104 / 0.349	9.20 / 24.70
Block 14 middle; t=0.9	9.10 / 21.40	14.90	—	0.122 / 0.410	5.10 / 15.30
Block 22 late; t=0.0	29.90 / 52.60	32.40	—	0.111 / 0.358	6.10 / 17.30
Block 22 late; t=0.1; y=null	31.80 / 54.90	34.30	—	0.112 / 0.354	6.20 / 17.90
Block 22 late; t=0.5	22.40 / 43.00	27.60	—	0.109 / 0.343	6.80 / 19.60
Block 22 late; t=0.9	7.30 / 18.40	12.90	—	0.126 / 0.446	4.40 / 14.10
Block 29 pre-output; t=0.0	28.80 / 50.90	30.90	—	0.114 / 0.358	6.00 / 17.20
Block 29 pre-output; t=0.1; y=null	30.40 / 53.00	32.90	—	0.110 / 0.343	6.10 / 17.50
Block 29 pre-output; t=0.5	22.70 / 43.40	26.90	—	0.116 / 0.357	6.70 / 18.90
Block 29 pre-output; t=0.9	7.40 / 18.20	12.60	—	0.129 / 0.379	4.80 / 14.50
Latent and generator controls					
Block 0; t=0.1; noised latent only	21.90 / 41.30	20.40	—	0.115 / 0.375	5.20 / 15.40
Block 0; t=0.1; zero latent	0.10 / 0.50	3.70	—	0.311 / 1.000	1.20 / 6.60
Block 0; t=0.1; shuffled latent	0.20 / 0.50	3.80	—	0.112 / 0.355	5.20 / 15.60
Block 14; t=0.1; untrained generator	11.70 / 25.50	12.90	—	0.121 / 0.418	3.00 / 10.40

The block-14, $t = 0.1$ baseline appears in Table 30. COCO uses mAP. The shuffled-latent NYUv2 intervention is a batch-size-one no-op.

C.11 Reference: VLM, MLLM, Multimodal, and Other Models

These category-specific records are separate from the unified LP/AP/FT comparison.

Table 35: VLM, MLLM, multimodal, and other reference models.

Model / setting	ImageNet-1K Top-1↑	COCO mAP↑	ADE20K mIoU↑	NYUv2 RMSE↓	SSv2 Top-1↑
CLIP	84.43	75.27	20.17	0.972	20.80
SigLIP 2 So400m	88.23	80.77	19.99	0.908	24.24
SigLIP 2 Giant	**89.00**	**81.90**	19.74	0.879	**27.29**
SAM 1	10.08	20.47	5.36	1.079	4.43
SAM 2.1	6.29	24.47	14.44	0.853	7.82
SAM 3	73.83	57.33	22.72	0.859	14.16
SAM 3.1	50.87	26.69	37.77	**0.753**	17.77
SAM 3D	71.84	6.14	**42.80**	0.799	8.99

Table 35: VLM/MLLM, multimodal, and other reference models. (continued)

Model / setting	**ImageNet-1K** Top-1↑	**COCO** mAP↑	**ADE20K** mIoU↑	**NYUv2** RMSE↓	**SSv2** Top-1↑
FLUX.2	9.07	1.24	17.87	0.901	1.66
Sapiens2	14.42	17.90	23.14	0.930	1.34
Molmo-7B-D	60.98	0.39	29.35	0.804	5.72
Molmo-7B-O	67.52	0.40	32.13	0.815	9.48
Qwen3-VL-8B	72.35	3.11	21.66	0.875	10.41
Cosmos 3 Super	29.36	30.84	7.41	1.144	1.44

C.12 Extended Basic Five Evaluation

Linear/minimal-head probing (LP), attentive probing (AP), and end-to-end fine-tuning (FT) are reported separately. Cells retain mean ± standard deviation and completed run counts n; a single run has no estimated dispersion.

Table 36: Unified Basic Five: linear/minimal-head probing (LP).

Model	**ImageNet-1K** Top-1 / Top-5	**COCO** Detection mAP	**ADE20K** mIoU / pAcc	**NYUv2** RMSE / AbsRel	**SSv2** Top-1 / Top-5
CLIP	79.56±0.01 95.84±0.01 $n = 3$	11.07±0.08 $n = 3$	33.74±0.04 67.85±0.05 $n = 3$	2.383±0.015 0.671±0.007 $n = 3$	20.48±1.04 46.84±1.04 $n = 3$
SigLIP 2 Giant	84.61±0.01 97.35±0.01 $n = 3$	35.51±0.06 $n = 3$	37.21±0.08 69.55±0.01 $n = 3$	2.023±0.032 0.528±0.014 $n = 3$	22.96±1.06 49.96±0.96 $n = 3$
C-RADIOv4-H	85.27±0.00 97.76±0.01 $n = 3$	**48.19±0.06** $n = 3$	46.99±0.13 **80.24±0.03** $n = 3$	1.461±0.020 0.341±0.007 $n = 3$	**25.85±1.58** **54.51±1.27** $n = 3$
Cosmos 3 Super	73.39±0.01 92.21±0.01 $n = 3$	36.72±0.02 $n = 3$	29.78±0.24 67.69±0.03 $n = 3$	0.862±0.004 0.231±0.001 $n = 3$	12.11±0.03 27.93±0.33 $n = 3$
VGGT-Ω	10.68±0.02 23.83±0.01 $n = 3$	0.50 $n = 3$	30.48±0.25 72.54±0.06 $n = 3$	0.678±0.009 0.162±0.003 $n = 3$	4.70±0.02 13.86±0.03 $n = 3$
RAEv2-K7	75.41±0.01 92.64±0.01 $n = 3$	38.91±0.12 $n = 3$	43.04±0.09 78.15±0.02 $n = 3$	0.662±0.039 0.205±0.021 $n = 3$	8.20±0.37 20.58±0.40 $n = 3$
V-JEPA 2.1 / ViT-g	57.61±0.03 82.74±0.01 $n = 3$	35.39±0.35 $n = 3$	36.10±0.12 74.89±0.05 $n = 3$	0.776±0.029 0.257±0.012 $n = 3$	21.51±0.22 45.45±0.19 $n = 3$
DINOv3-7B	**86.54±0.01** **97.80±0.00** $n = 3$	40.77±0.04 $n = 3$	**47.14±0.07** 79.78±0.05 $n = 3$	**0.577±0.001** **0.159±0.000** $n = 3$	14.90±0.23 34.52±0.21 $n = 3$

Table 37: Unified Basic Five: attentive probing (AP).

Model	ImageNet-1K Top-1 / Top-5	COCO Detection mAP	ADE20K mIoU / pAcc	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
CLIP	84.38±0.11 97.31±0.05 $n = 3$	13.87±0.04 $n = 3$	38.82±0.11 75.26±0.08 $n = 3$	0.834±0.005 0.226±0.001 $n = 3$	33.32±0.53 66.33±0.14 $n = 3$
SigLIP 2 Giant	87.36±0.03 **98.14±0.00** $n = 3$	38.67±0.10 $n = 3$	44.09±0.11 78.01±0.11 $n = 3$	0.692±0.004 0.179±0.001 $n = 3$	40.90±0.57 73.70±0.64 $n = 3$
C-RADIOv4-H	86.49±0.13 97.82±0.04 $n = 3$	**50.04±0.10** $n = 3$	**48.10±0.18** **81.26±0.04** $n = 3$	0.590±0.004 0.146±0.002 $n = 3$	43.58±0.24 75.44±0.05 $n = 3$
Cosmos 3 Super	82.76±0.11 96.47±0.06 $n = 3$	38.06±0.15 $n = 3$	34.00±0.39 72.21±0.13 $n = 3$	0.811±0.002 0.212±0.001 $n = 3$	35.61±0.25 68.20±0.17 $n = 3$
VGGT-Ω	81.62±0.07 95.91±0.05 $n = 3$	11.63±0.43 $n = 3$	27.18±0.05 73.01±0.07 $n = 3$	0.627±0.005 0.157±0.003 $n = 3$	44.57±0.25 76.53±0.28 $n = 3$
RAEv2-K7	86.26±0.11 97.58±0.04 $n = 3$	41.31±0.51 $n = 3$	45.19±0.39 80.25±0.08 $n = 3$	0.476±0.002 0.130±0.001 $n = 3$	40.69±0.08 72.59±0.20 $n = 3$
V-JEPA 2.1 / ViT-g	81.39±0.13 95.66±0.02 $n = 3$	35.65±0.28 $n = 3$	0.22±0.02 13.83±1.35 $n = 3$	0.473±0.001 0.129±0.001 $n = 3$	**69.21±0.17** **92.62±0.12** $n = 3$
DINOv3-7B	**87.82±0.01** 97.89±0.03 $n = 3$	43.78±0.07 $n = 3$	47.91±0.07 80.56±0.04 $n = 3$	**0.423±0.001** **0.114±0.000** $n = 3$	47.43±0.34 78.81±0.13 $n = 3$

Table 38: Unified Basic Five: end-to-end fine-tuning (FT). Due to the substantial computational cost, some FT runs could not be completed. Only configurations with completed evaluations are reported.

Model	ImageNet-1K Top-1 / Top-5	COCO Detection mAP	ADE20K mIoU / pAcc	NYUv2 RMSE / AbsRel	SSv2 Top-1 / Top-5
CLIP	84.90±0.04 96.71±0.07 $n = 3$	25.72±0.12 $n = 3$	42.88±0.24 78.33±0.18 $n = 3$	1.200±0.157 0.270±0.024 $n = 3$	38.92±0.31 70.11±0.22 $n = 3$
SigLIP 2 Giant	87.00±0.09 97.68±0.09 $n = 3$	**47.28±0.13** $n = 3$	46.23±0.27 78.86±0.07 $n = 3$	1.981±0.081 0.545±0.034 $n = 3$	43.71±0.09 76.34±0.29 $n = 3$
C-RADIOv4-H	83.76±0.05 96.72±0.08 $n = 3$	45.83±0.12 $n = 3$	**50.08±0.35** **82.44±0.10** $n = 3$	0.492±0.002 0.117±0.000 $n = 3$	46.67±1.70 76.21±1.69 $n = 3$
Cosmos 3 Super	85.02±0.08 97.25±0.05 $n = 3$	46.91±0.03 $n = 3$	38.55±0.13 75.76±0.05 $n = 3$	0.665±0.005 0.161±0.001 $n = 3$	35.72±0.31 67.43±0.25 $n = 3$
VGGT-Ω	87.05±0.12 98.28±0.06 $n = 3$	6.86±0.52 $n = 3$	44.48±0.51 80.86±0.13 $n = 3$	0.467±0.001 0.105±0.000 $n = 3$	63.17 89.77 $n = 1$
RAEv2-K7	**88.07±0.04** **98.50±0.02** $n = 3$	42.93±0.46 $n = 3$	42.98±0.14 79.13±0.05 $n = 3$	0.476±0.002 0.129±0.001 $n = 3$	45.04±0.32 76.68±0.19 $n = 3$
V-JEPA 2.1 / ViT-g	—	0.07±0.12 $n = 3$	35.88±0.20 75.48±0.16 $n = 3$	0.426±0.001 0.112±0.000 $n = 3$	**76.45±0.07** **95.81±0.07** $n = 3$
DINOv3-7B	—	—	48.42±0.03 80.87±0.01 $n = 3$	**0.392±0.001** **0.103±0.000** $n = 3$	—

C.13 Extended Evaluation under Unified LP and AP Protocols

Tables 39 and 40 report extended evaluation under unified LP and AP protocols, respectively, for five model configurations across 45 datasets. We include datasets with recorded scores for all five configurations under both protocols and retain the task-specific metrics and their directions.

Table 39: Extended evaluation on 45 datasets under the unified LP protocol. These results correspond to Figure 2.

Dataset	Metric	DINOv3 7B	VGGT-Ω	RAEv2 K7	V-JEPA 2.1	SigLIP 2 Giant
Action40 (Yao et al., 2011)	Top-1 ↑	94.87	18.67	74.11	69.97	92.86
Stanford Cars (Krause et al., 2013)	Top-1 ↑	87.49	0.99	15.16	5.02	89.17
CLUE (egg) (Shinoda et al., 2025)	Top-1 ↑	39.75	10.60	29.78	26.62	26.00
CLUE (feather) (Shinoda et al., 2025)	Top-1 ↑	42.79	11.91	25.29	19.71	25.89
CLUE (footprint) (Shinoda et al., 2025)	Top-1 ↑	26.03	6.73	15.41	12.81	10.33
CLUE (skulls) (Shinoda et al., 2025)	Top-1 ↑	14.10	2.48	5.12	5.36	6.10
CLUE (stools) (Shinoda et al., 2025)	Top-1 ↑	38.66	7.42	25.09	19.07	16.30
CUB-200-2011 (Wah et al., 2011)	Top-1 ↑	88.35	3.69	25.72	12.39	79.32
DTD (Cimpoi et al., 2014)	Top-1 ↑	80.37	17.07	70.85	53.40	70.00
Oxford Flowers-102 (Nilsback & Zisserman, 2008)	Top-1 ↑	99.69	5.66	98.24	90.01	97.35
Food-101 (Bossard et al., 2014)	Top-1 ↑	96.22	12.46	89.57	62.00	94.79
ImageNet-10% (Russakovsky et al., 2015)	Top-1 ↑	84.71	4.92	65.65	47.63	81.36
ImageNet-1% (Russakovsky et al., 2015)	Top-1 ↑	81.36	2.40	51.95	33.69	71.62
IP102 (Wu et al., 2019)	Top-1 ↑	69.88	16.49	47.82	38.99	57.87
MIT Indoor-67 (Quattoni & Torralba, 2009)	Top-1 ↑	91.12	14.48	84.85	71.12	90.37
Oxford-IIIT Pets (Parkhi et al., 2012)	Top-1 ↑	96.57	11.88	86.84	50.86	95.20
SUN397 (Xiao et al., 2010)	Top-1 ↑	78.37	5.93	64.92	52.93	76.40
FGVC-Aircraft (Maji et al., 2013)	Top-1 ↑	64.51	4.86	33.81	10.26	69.73
CIFAR-10 (Krizhevsky, 2009)	Top-1 ↑	99.46	71.13	98.78	97.25	97.80
CIFAR-100 (Krizhevsky, 2009)	Top-1 ↑	94.48	16.16	89.18	82.00	84.28
KMNIST (Clanuwat et al., 2018)	Top-1 ↑	66.58	30.22	62.13	56.56	58.15
Kuzushiji-Kanji (Clanuwat et al., 2018)	Top-1 ↑	2.72	0.29	0.90	1.31	1.93
MNIST (LeCun et al., 1998)	Top-1 ↑	96.82	59.51	91.77	88.27	96.91
OG Omniglot15 (Lake et al., 2015)	Top-1 ↑	36.00	0.44	21.26	3.76	34.87
HMDB51 (Kuehne et al., 2011)	Top-1 ↑	54.90	28.56	35.95	42.68	65.42
UCF101 (Soomro et al., 2012)	Top-1 ↑	89.72	50.99	51.04	31.62	93.29
COCO 2017 (Lin et al., 2014b)	box AP ↑	40.79	19.42	38.93	35.20	35.49
LIVECell (Edlund et al., 2021)	mask AP ↑	0.27	0.08	11.80	10.72	7.22
LVIS v1 (Gupta et al., 2019)	mask AP ↑	6.08	3.44	16.63	12.42	13.85
ADE20K (Zhou et al., 2017)	mIoU ↑	44.73	30.95	42.64	33.08	34.45
PASCAL VOC 2012 (Everingham et al., 2012)	mIoU ↑	80.73	58.66	79.41	70.80	65.42
SpaceNet 2 Vegas (SpaceNet on Amazon Web Services (AWS), 2018)	mIoU ↑	72.40	67.46	71.31	67.89	66.02
Mapillary Vistas (Neuhold et al., 2017)	mIoU ↑	32.88	24.97	30.33	24.12	22.43
BDD100K (Yu et al., 2020)	mIoU ↑	47.76	39.24	46.37	39.20	38.55
PASCAL Context (Mottaghi et al., 2014)	mIoU ↑	59.10	41.13	57.50	51.06	43.73
SUN RGB-D (Song et al., 2015)	mIoU ↑	52.06	42.27	47.89	43.24	38.70
MPII Human Pose (Andriluka et al., 2014)	PCKh ↑	50.09	46.71	48.14	51.27	42.36
CrowdPose (Li et al., 2019)	keypoint AP ↑	6.78	8.50	6.29	7.57	6.56
DAVIS 2017 (Pont-Tuset et al., 2017)	F ↑	9.41	1.36	18.72	13.37	18.86
MPI Sintel (Butler et al., 2012)	EPE ↓	3.6803	3.5381	3.5609	3.5584	3.4817
Middlebury Flow (Baker et al., 2011)	EPE ↓	5.5871	5.5772	5.6099	5.5895	5.5846
7-Scenes (Shotton et al., 2013)	t_{med} (m) ↓	0.2524	0.3314	0.3104	0.3758	0.4082
Cambridge Landmarks (Kendall et al., 2015)	t_{med} (m) ↓	27.7128	17.7516	29.2645	32.3814	13.8449
3DSRBench (Ma et al., 2025)	Accuracy ↑	33.49	33.97	35.89	35.89	33.49
NYU Depth V2 (Silberman et al., 2012)	RMSE ↓	0.5769	2.4173	0.6273	0.7650	2.1626

Arrows indicate the better direction. t_{med}: median translation error.

Table 40: Extended evaluation on 45 datasets under the unified AP protocol. These results correspond to Figure 2.

Dataset	Metric	DINOv3 7B	VGGT-Ω	RAEv2 K7	V-JEPA 2.1	SigLIP 2 Giant
Action40 (Yao et al., 2011)	Top-1 ↑	96.49	90.84	94.52	90.53	96.04
Stanford Cars (Krause et al., 2013)	Top-1 ↑	94.29	85.69	92.65	85.47	95.06
CLUE (egg) (Shinoda et al., 2025)	Top-1 ↑	52.90	45.81	48.68	44.41	45.49
CLUE (feather) (Shinoda et al., 2025)	Top-1 ↑	63.35	55.50	61.59	54.46	55.67
CLUE (footprint) (Shinoda et al., 2025)	Top-1 ↑	31.19	22.40	24.47	20.03	20.21
CLUE (skulls) (Shinoda et al., 2025)	Top-1 ↑	29.68	14.81	21.97	14.52	18.38
CLUE (stools) (Shinoda et al., 2025)	Top-1 ↑	43.39	33.47	39.68	34.32	35.50
CUB-200-2011 (Wah et al., 2011)	Top-1 ↑	87.66	78.29	84.78	75.15	86.81
DTD (Cimpoi et al., 2014)	Top-1 ↑	81.01	68.40	78.83	71.60	81.54
Oxford Flowers-102 (Nilsback & Zisserman, 2008)	Top-1 ↑	99.69	92.50	99.38	97.90	99.19
Food-101 (Bossard et al., 2014)	Top-1 ↑	95.94	90.34	94.86	88.13	95.03
ImageNet-10% (Russakovsky et al., 2015)	Top-1 ↑	83.33	71.66	78.51	67.90	82.11
ImageNet-1% (Russakovsky et al., 2015)	Top-1 ↑	77.49	52.48	68.09	52.39	73.40
IP102 (Wu et al., 2019)	Top-1 ↑	76.26	70.55	75.71	71.35	74.65
MIT Indoor-67 (Quattoni & Torralba, 2009)	Top-1 ↑	89.55	82.31	88.28	83.96	91.27
Oxford-IIIT Pets (Parkhi et al., 2012)	Top-1 ↑	96.13	89.10	94.30	89.92	96.18
SUN397 (Xiao et al., 2010)	Top-1 ↑	76.08	65.66	74.55	66.68	78.50
FGVC-Aircraft (Maji et al., 2013)	Top-1 ↑	90.85	67.54	85.96	79.84	82.21
CIFAR-10 (Krizhevsky, 2009)	Top-1 ↑	99.59	97.46	99.23	98.49	98.65
CIFAR-100 (Krizhevsky, 2009)	Top-1 ↑	94.96	83.81	92.23	89.40	89.79
KMNIST (Clanuwat et al., 2018)	Top-1 ↑	97.52	92.98	97.58	97.75	96.57
Kuzushiji-Kanji (Clanuwat et al., 2018)	Top-1 ↑	66.25	61.64	71.08	67.63	69.58
MNIST (LeCun et al., 1998)	Top-1 ↑	98.87	98.48	99.03	98.94	99.04
OG Omniglot15 (Lake et al., 2015)	Top-1 ↑	80.38	77.84	80.44	78.02	80.86
HMDB51 (Kuehne et al., 2011)	Top-1 ↑	65.88	28.89	61.37	78.82	67.91
UCF101 (Soomro et al., 2012)	Top-1 ↑	92.47	87.76	88.00	95.35	90.75
COCO 2017 (Lin et al., 2014b)	box AP ↑	31.33	21.54	41.13	35.45	38.68
LIVECell (Edlund et al., 2021)	mask AP ↑	0.33	0.10	10.14	9.77	6.57
LVIS v1 (Gupta et al., 2019)	mask AP ↑	6.55	3.45	16.13	11.32	14.15
ADE20K (Zhou et al., 2017)	mIoU ↑	42.59	22.34	42.95	30.41	40.73
PASCAL VOC 2012 (Everingham et al., 2012)	mIoU ↑	80.28	46.20	78.09	69.26	74.30
SpaceNet 2 Vegas (SpaceNet on Amazon Web Services (AWS), 2018)	mIoU ↑	74.51	71.17	74.84	65.44	70.46
Mapillary Vistas (Neuhold et al., 2017)	mIoU ↑	33.63	26.47	30.94	25.84	29.14
BDD100K (Yu et al., 2020)	mIoU ↑	46.97	38.86	47.45	37.73	43.64
PASCAL Context (Mottaghi et al., 2014)	mIoU ↑	57.26	37.62	57.10	49.12	53.70
SUN RGB-D (Song et al., 2015)	mIoU ↑	46.16	38.69	48.06	42.00	45.01
MPII Human Pose (Andriluka et al., 2014)	PCKh ↑	79.37	69.69	80.46	76.90	75.72
CrowdPose (Li et al., 2019)	keypoint AP ↑	47.28	37.83	51.01	46.28	46.41
DAVIS 2017 (Pont-Tuset et al., 2017)	F ↑	1.28	1.28	1.28	1.28	1.28
MPI Sintel (Butler et al., 2012)	EPE ↓	5.0914	3.5704	4.5938	4.6981	3.4818
Middlebury Flow (Baker et al., 2011)	EPE ↓	5.5798	5.6013	5.5469	5.6292	5.5697
7-Scenes (Shotton et al., 2013)	t_{med} (m) ↓	0.1828	0.2665	0.2109	0.2104	0.2564
Cambridge Landmarks (Kendall et al., 2015)	t_{med} (m) ↓	2.7433	10.3799	4.1236	5.7909	5.7969
3DSRBench (Ma et al., 2025)	Accuracy ↑	28.71	33.49	34.21	24.88	33.97
NYU Depth V2 (Silberman et al., 2012)	RMSE ↓	0.4224	0.627	0.4760	0.4724	0.692

Arrows indicate the better direction. t_{med}: median translation error.

C.14 Extended Evaluation: Protocol-Specific Reference Records

Extended evaluation with six model configurations at each protocol. The two panels retain the recorded checkpoint, interface, and task protocol and are supplementary references, not a unified evaluation shown in Figure 2.

Table 41: Extended evaluation with six model configurations under their respective protocols.

Dataset / evaluation entry	DINOv3	DINOv3 7B	VGGT-Ω	Cosmos 3 Super	RAEv2 K7	V-JEPA 2.1
Action40 (Yao et al., 2011) ↑	72.34	95.07	20.75	61.58	77.48	69.75
ADE20K (Zhou et al., 2017) ↑	24.96	31.73	13.01	7.41	45.84	26.96
Cars (Krause et al., 2013) ↑	55.41	91.99	2.27	0.846	68.97	7.69
CIFAR-10 (Krizhevsky, 2009) ↑	72.42	99.60	63.85	67.45	98.66	98.61
CIFAR-100 (Krizhevsky, 2009) ↑	47.69	95.52	26.68	38.34	90.65	87.26
CLUE (egg) (Shinoda et al., 2025) ↑	22.72	49.30	12.28	6.32	31.52	28.51
CLUE (feather) (Shinoda et al., 2025) ↑	29.74	56.87	11.88	7.87	29.20	21.33
CLUE (footprint) (Shinoda et al., 2025) ↑	15.30	36.52	7.07	4.04	22.07	16.17
CLUE (skulls) (Shinoda et al., 2025) ↑	7.68	21.72	4.09	1.36	9.11	9.26
CLUE (stools) (Shinoda et al., 2025) ↑	16.72	44.70	6.76	2.27	31.36	22.32
COCO2017 (Lin et al., 2014b) ↑	59.29	82.40	20.99	30.84	76.72	62.16
CUB200 (Wah et al., 2011) ↑	25.54	89.98	4.66	8.59	21.34	10.59
DAVIS2017 (Pont-Tuset et al., 2017) ↑	10.91	10.92	6.17	7.83	12.36	12.51
DomainNet (Peng et al., 2019) ↑	63.64	84.14	8.13	2.23	75.28	51.96
DomainNet: infograph (Peng et al., 2019) ↑	31.02	66.58	4.84	1.75	40.42	20.91
DomainNet: quickdraw (Peng et al., 2019) ↑	60.75	70.57	18.15	11.28	59.37	41.65
DomainNet: real (Peng et al., 2019) ↑	81.71	91.75	27.12	14.45	88.26	78.25
DomainNet: sketch (Peng et al., 2019) ↑	63.64	84.14	8.13	2.23	75.28	51.96
DTD (Cimpoi et al., 2014) ↑	55.87	80.55	22.32	34.91	74.82	57.11
Endoscapes (Murali et al., 2023) ↑	0.371	0.431	0.182	0.204	0.362	0.362
ETH3D (Schöps et al., 2017) ↑	77.78	100.00	55.56	11.11	66.67	77.78
Flowers-102 (Nilsback & Zisserman, 2008) ↑	88.20	99.69	19.85	56.31	99.59	97.16
Food-101 (Bossard et al., 2014) ↑	71.40	96.76	18.92	42.97	90.55	66.43
HMDB51 (Kuehne et al., 2011) ↑	32.89	77.74	10.92	8.10	24.62	46.10
ImageNet-100 (Russakovsky et al., 2015) ↑	88.65	97.54	42.15	58.24	94.84	88.13
ImageNet-1percent (Russakovsky et al., 2015) ↑	42.51	82.79	4.15	4.79	61.15	39.19
ImageNet-10percent (Russakovsky et al., 2015) ↑	60.83	86.34	11.29	25.91	76.14	56.85
ImageNet-1K (Russakovsky et al., 2015) ↑	69.57	88.12	31.40	29.36	84.99	70.18
ImageNet-A (Hendrycks et al., 2021b) ↑	25.72	97.78	7.56	2.62	29.39	31.84
ImageNet-C (Hendrycks & Dietterich, 2019) ↑	21.65	97.37	19.26	23.61	79.12	66.95
ImageNet-R (Hendrycks et al., 2021a) ↑	69.88	98.52	4.04	1.53	76.63	50.22
ImageNet-Sketch (Wang et al., 2019) ↑	77.00	96.05	6.79	2.73	77.82	54.95
Industrial iSeg (Li & Wong, 2024) ↑	0.319	0.344	0.307	0.191	0.329	0.322
IP102 (Wu et al., 2019) ↑	50.00	75.32	18.11	12.76	55.74	42.63
Kinetics-400 (Kay et al., 2017) ↑	36.46	75.73	13.73	5.29	60.51	53.83
KITTI (Geiger et al., 2013) ↓	9.49	9.11	9.37	9.83	9.48	9.24
KMNIST (Clanuwat et al., 2018) ↑	72.11	78.87	41.50	48.61	63.91	56.02
Kuzushiji (Clanuwat et al., 2018) ↑	17.54	15.19	0.994	0.180	6.18	5.14
LiveCell (Edlund et al., 2021) ↑	0.575	0.658	0.693	0.529	0.674	0.702
MegaDepth (Li & Snavely, 2018) ↓	2.89	2.66	3.49	3.71	3.20	2.98
MIT Indoor (Quattoni & Torralba, 2009) ↑	70.60	91.04	20.27	11.39	87.76	75.17
MNIST (LeCun et al., 1998) ↑	95.87	98.39	65.33	67.36	94.24	85.79
NYUv2 (Silberman et al., 2012) ↓	0.889	0.641	0.734	1.144	0.056	0.755
ObjectNet (Barbu et al., 2019) ↑	74.85	95.39	26.30	22.03	57.03	45.91

Table 41: Extended evaluation with six model configurations under their respective protocols. (continued)

Dataset / evaluation entry	DINOv3	DINOv3 7B	VGGT-Ω	Cosmos 3 Super	RAEv2 K7	V-JEPA 2.1
OG (Omniglot15) (Lake et al., 2015) ↑	65.61	75.76	3.96	5.38	46.73	14.08
Oxford Pets (Parkhi et al., 2012) ↑	73.74	96.34	9.92	19.02	83.12	43.39
PASCAL VOC2012 (Everingham et al., 2012) ↑	60.43	84.94	31.65	31.59	69.17	88.24
Places30 (Zhou et al., 2018) ↑	77.21	82.69	57.78	63.20	81.42	78.81
Places365 (Zhou et al., 2018) ↑	81.01	88.15	60.99	62.70	87.05	82.87
SevenScenes (Shotton et al., 2013) ↑	59.26	93.52	27.78	27.78	59.72	65.05
SpaceNet2 Vegas (SpaceNet on Amazon Web Services (AWS), 2018) ↑	0.654	0.724	0.725	0.567	0.729	0.735
SSv2 (Goyal et al., 2017) ↑	17.85	29.81	8.00	1.44	1.48	41.14
SUN397 (Xiao et al., 2010) ↑	60.08	81.66	15.40	4.49	75.04	63.29
UCF101 (Soomro et al., 2012) ↑	49.50	92.34	16.86	2.17	71.41	70.48

Lower is better for KITTI, MegaDepth, and NYUv2. Other entries retain their recorded higher-is-better metrics.

C.15 Reference: Frontier Models

We compare DINOv3-7B with linear probing (LP) and attentive probing (AP) against GPT-5.6 Sol and GPT-6 Astra as a supplementary frontier model reference. Each dataset uses the same fixed subset of 500 evaluation samples across all four entries. Task interfaces are adapted to the prompted GPT systems, and predictions from the already-trained DINOv3 heads are mapped to the same targets without retraining. ImageNet-1K and SSv2 use classification accuracy; COCO uses image-level category-presence Micro-F1; ADE20K uses semantic classification accuracy at 32 query points per image; and NYUv2 uses relative-depth accuracy over 20 point pairs per image. These measures are distinct from the native Basic Five detection, segmentation, and metric-depth measures. The detailed evaluation protocols are provided in Appendix E.

Table 42: Frontier-model reference comparison on matched 500-sample subsets. All values are percentages; higher is better. Blue shading and bold indicate the highest observed score in each column, not statistical significance. These results are excluded from the native Basic Five LP/AP/FT rankings.

Model / interface	ImageNet-1K Top-1	COCO Micro-F1	ADE20K Point accuracy	NYUv2 Relative-depth accuracy	SSv2 Top-1
GPT-6 Astra	94.40	86.02	66.96	80.02	**70.20**
GPT-5.6 Sol	**95.20**	**87.59**	58.91	68.23	56.80
DINOv3-7B LP	85.80	70.77	79.99	**80.30**	15.60
DINOv3-7B AP	87.40	70.50	**80.98**	79.00	43.40

Table 42 reveals complementary strengths. DINOv3-7B exceeds both GPT systems on ADE20K semantic-point accuracy, while its LP relative-depth accuracy is close to GPT-6 Astra's. Both GPT systems achieve higher scores on ImageNet-1K, COCO category presence, and SSv2. Matching samples, targets, and scoring rules does not equate training data, computational budgets, or evaluation interfaces: the GPT entries are prompted end-to-end systems, whereas DINOv3 uses a frozen visual backbone with task-trained heads. The comparison therefore provides context for vision-native representations rather than isolating representation quality or establishing an overall model ranking.

D VISUALIZATION AND ANALYSIS OF RESULTS

D.1 PAIR-PLOT OF THE BASIC FIVE

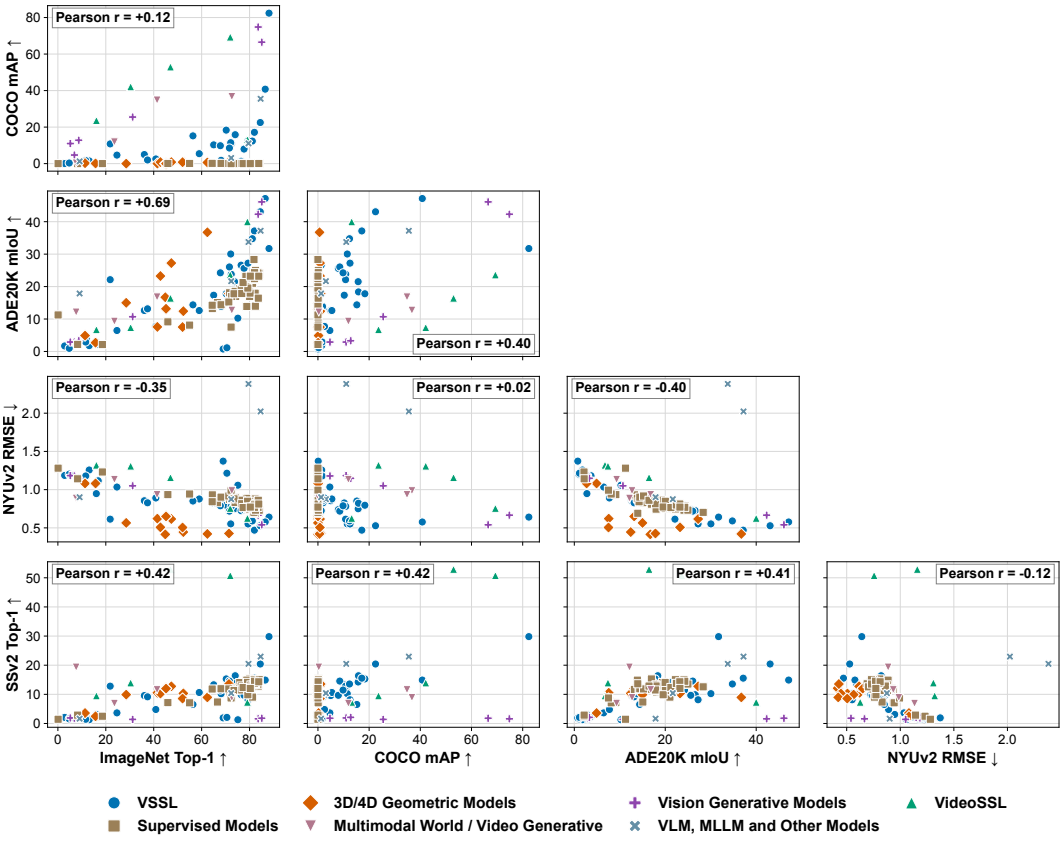


Figure 13: Pairwise relationships among the Basic Five under ASIS. Each point represents an evaluated model configuration with complete scores for ImageNet-1K classification, COCO detection, ADE20K segmentation, NYUv2 depth estimation, and SSv2 video classification. Colors and marker shapes indicate model categories, and each panel reports Pearson's r on the native metric scales. Associations vary across task pairs, with the strongest positive relationship observed between ImageNet-1K and ADE20K ($r = 0.69$). Because NYUv2 is measured by RMSE, lower values are better and the interpretation of its correlation sign is reversed relative to the other metrics. These associations are descriptive and do not imply matched training conditions or causal transfer. Wan2.2 is excluded because its ADE20K result is unavailable, while AlayaWorld v1.1 is omitted from the visualization to preserve readable axis scales.

D.2 TRAINING PROGRESS OF THE BASIC FIVE ON CTRL

D.3 DINOV3 SCALING ON THE BASIC FIVE

Figure 15 compares the ViT-B, ViT-L, ViT-H, and 7B variants of DINOv3 on the Basic Five. We report ImageNet-1K and SSv2 Top-1 accuracy, COCO AP, ADE20K mIoU, and NYUv2 RMSE, retaining the native scale of each metric. The figure summarizes three-seed results for each configuration; model settings and evaluation protocols are described in Appendices B and E.

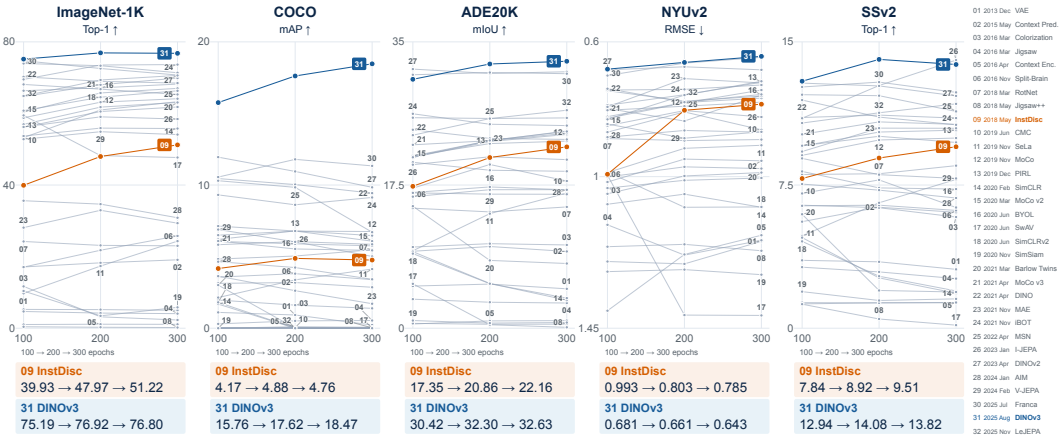


Figure 14: Training progress of 32 VSSL methods on the Basic Five under CTRL. Each method uses a ViT-B/16 encoder pre-trained on unlabeled ImageNet-1K under common CTRL pre-training settings, while retaining method-specific mechanisms. Downstream performance is evaluated at the 100-, 200-, and 300-epoch checkpoints of the same training trajectory for each method, rather than from independently trained runs. Lines connect the three recorded checkpoints, with numbered labels referring to the method roster at right. Instance Discrimination (InstDisc) and DINOv3 are highlighted in orange and blue, respectively; the lower callouts report their scores in 100 → 200 → 300 epoch order. Higher is better for ImageNet-1K, COCO, ADE20K, and SSv2. Lower is better for NYUv2 RMSE, whose vertical axis is reversed so that upward movement indicates improvement in every panel. Equal epoch counts do not imply equal computation, and cross-method comparisons do not isolate the learning objective alone because method-specific mechanisms and evaluation settings remain.

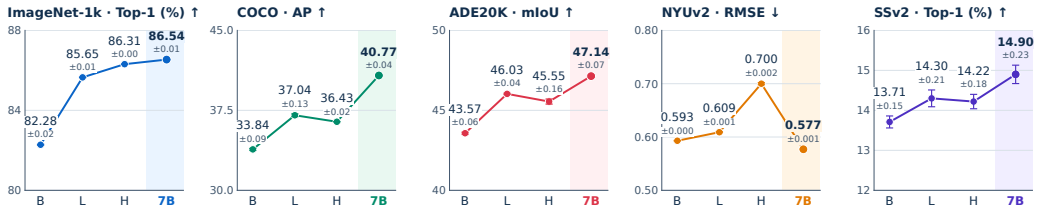


Figure 15: DINOv3 scaling on the Basic Five. Points show means and error bars show the reported variability across three seeds. Shading highlights 7B, and bold values indicate the best mean per task. Lower NYUv2 RMSE is better; higher is better for the other metrics.

E EVALUATION PROTOCOLS

We summarize the evaluation scope, protocol families, and task definitions here. Companion Markdown documents and JSON registries are provided in the accompanying supplementary archive rather than reproduced in full in this PDF. The eight initial-comparison documents are located under `docs/initial_protocols/`; the unified LP/AP/FT specifications and frontier-system prompts are located under `docs/submission_protocols/`. The filenames below identify these documents independently of the archive name or hosting location. Checkpoint and component identities appear in Appendix B, and scores, repeat counts, and evaluation coverage in Appendix C.

The initial comparisons and later unified evaluations use distinct protocols. Historical results retain their original settings; reconstructed specifications are identified separately and are not treated as execution records.

E.1 SCOPE AND REPLICATION

The unified protocols apply to representative-model comparisons, separately from historical ASIS/CTRL and component-specific experiments. Reporting conventions are documented in `00_scope_and_replication.md`. Basic Five results report the actual number of completed runs, with three-seed and single-seed evaluations distinguished in Appendix C.

Multiple valid runs are summarized by their mean, sample standard deviation, and repetition count. Single-run results have no estimated standard deviation. Incomplete three-seed coverage is reported with its actual count rather than presented as a completed three-seed experiment. Extended evaluations use seed 0.

Technically valid low or zero scores remain results; unavailable, incomplete, or technically invalid evaluations are not measured zeros. Exclusions require a technical reason rather than a score threshold. Protocol catalogs and scheduling instructions specify intended scope, not completed experiments.

E.2 ASIS AND CTRL COMPARISONS

ASIS evaluates public weights or reproductions following the original architecture and learning procedure. Its initial checkpoint scope and downstream documentation are provided in `INITIAL_STEP1_v1.md`. The corresponding CTRL documentation is provided in `INITIAL_STEP2_v1.md`. The Step identifiers in these filenames are legacy project identifiers, not subsection numbers in this appendix.

CTRL uses ViT-B/16 on unlabeled ImageNet-1K for 300 epochs, with checkpoints at epochs 100, 200, and 300 from the same trajectory. Its shared pre-training recipe uses 224×224 inputs, seed 42, effective batch 1,024, AdamW with peak learning rate 6×10^{-4} and weight decay 0.05, cosine decay, and 10 warmup epochs, retaining essential method-specific mechanisms. The reconstructed final-checkpoint downstream specification uses the epoch-300 checkpoint; intermediate-checkpoint comparisons identify their checkpoint explicitly.

CTRL aligns key pre-training conditions, not every downstream predictor or computational cost. Historical downstream results retain their recorded splits, predictors, schedules, and checkpoint-selection rules and are distinguished from the later unified Basic Five evaluations.

E.3 INITIAL CATEGORY-SPECIFIC COMPARISONS

Six companion documents describe the initial category-specific comparisons, distinguishing recorded settings from reconstructed specifications.

- **3D foundation models (3DFM).** `INITIAL_STEP3_3DFM_v1.md` documents the initial frozen-feature comparisons using encoder or aggregator patch representations from 3D foundation models, including VGGT-Ω.
- **4D foundation models (4DFM).** `INITIAL_STEP3_4DFM_v1.md` documents the initial comparisons of dynamic-scene and temporally structured foundation models, including V-DPM.

- **Vision generative models.** `INITIAL_STEP3_VISION_GENERATIVE_v1.md` documents the initial RAE/MAGE comparison, including the historical RAEv2 aggregation of class tokens from the last seven layers. This configuration is distinct from final-layer-only extraction.
- **Video self-supervised models (VideoSSL).** `INITIAL_STEP3_VIDEO_SSL_v1.md` documents the initial video-representation comparison. Historical seed information is unavailable for some runs; these results are not treated as replicated estimates.
- **Video generative and world models.** `INITIAL_STEP3_VIDEO_WORLD_MODELS_v1.md` concerns the initial Cosmos 3 Super vision-encoder adapter, rather than all models in this category. Its representation is the 1152-dimensional pre-merger `last_hidden_state`, with the documented historical preprocessing. It is not the later 5120-dimensional merger-output representation or an evaluation of the MoT, DiT, or VAE components.
- **Vision–language models (VLM).** `INITIAL_STEP3_VLM_v1.md` documents the initial SigLIP 2 So400m/14 comparison at 384-pixel resolution, using MAP-pooled global features and spatial vision features. This scope is distinct from later giant-model comparisons.

Each result retains its recorded checkpoint, feature interface, preprocessing, task definition, and evaluation protocol.

E.4 Basic Five: Tasks and Evaluation Tracks

The unified tracks use official dataset splits and fixed task definitions. ImageNet-1K uses the ILSVRC2012 train/validation split and Top-1 accuracy. COCO uses train2017/val2017 and bounding-box AP over IoU thresholds 0.50–0.95, with Faster R-CNN/FPN and 800–1,333-pixel resizing. ADE20K uses 20,210 training and 2,000 validation images, 150 classes, and mIoU. NYUv2 uses the official 795-training/654-test split, valid depths of 0.1–10 m, and RMSE in meters without median alignment. SSv2 uses the official train/validation split, 174 classes, 16 temporal segments, and Top-1 accuracy, with temporally consistent crops and no horizontal flipping. Non-detection tasks use 224-pixel inputs under the specified preprocessing rules.

The scheduled training lengths are 100 epochs for ImageNet-1K, 12 epochs for COCO, 20 epochs for ADE20K, 30 epochs for NYUv2, and 50 epochs for SSv2. The unified results use the final scheduled checkpoint rather than substituting a best-validation checkpoint. Secondary metrics and task-specific optimization settings are specified in the companion documents.

Linear probing or frozen-feature probing (LP). The backbone remains frozen in evaluation mode; only task-specific predictors are trained. These are linear classifiers for image/video classification, Faster R-CNN/FPN for detection, and spatial 1×1 heads for segmentation and depth. Thus, LP is a track identifier, not a claim that every predictor is linear. See `BASIC5_LINEAR.md`.

Attentive probing (AP). AP adds an attention reader or spatial adapter before the task head while keeping the backbone frozen. Video attention receives ordered temporal or spatiotemporal tokens. The task definitions, dataset splits, and scheduled training lengths remain aligned with the LP specification. See `BASIC5_ATTENTIVE.md`.

Fine-tuning (FT). FT independently initializes each task from the designated checkpoint and updates the backbone and baseline task head, without the AP reader. It is not a continuation from an LP or AP checkpoint. See `BASIC5_FINETUNE.md`. Optimization and augmentation can differ across tracks, so their differences do not isolate predictor architecture or trainable scope alone.

E.5 Extended Evaluation

The extended suite covers recognition, dense prediction, video, geometry, correspondence, and question-conditioned tasks. `EXTEND_LINEAR.md` specifies the base task families and reporting rules; `EXTEND_ATTENTIVE.md` and `EXTEND_FINETUNE.md` define the corresponding AP and FT tracks.

The accompanying registries are `EXTEND_LINEAR_v1.json`, `EXTEND_ATTENTIVE_v1.json`, and `EXTEND_FINETUNE_v1.json`. They provide the base dataset configuration and track-specific inheritance or overrides. The package `README.md` describes the supplied version and its interpretation.

The Markdown documents and JSON registries specify the evaluation protocols. Reproducing individual results additionally requires the corresponding run configurations, dataset bindings, and implementation versions; protocol specifications alone are not execution records.

A Basic Five result is reused only when its checkpoint, individual seed, split, features, predictor, optimization, preprocessing, metric, and evaluation procedure match. For seed-0 extended evaluation, reuse means the individual seed-0 result, not a three-seed mean. LP/AP/FT remain separate, and each summary identifies its evaluated entries, missing results, and denominator rather than treating the catalog size as completed coverage.

E.6 Frontier-System Reference

The manuscript-level frontier-system comparison specifies GPT-5.6 Sol, GPT-6 Astra, and the already-trained DINOv3-7B LP/AP readouts on matched, fixed subsets of 500 held-out samples per dataset. The subset seed identifies sample selection, not closed-system decoding. ImageNet-1K and SSv2 use Top-1 accuracy; COCO uses category-presence Micro F1, ADE20K semantic-point accuracy at 32 points per image, and NYUv2 relative-depth accuracy over 20 coordinate pairs per image. SSv2 inputs contain 16 ordered frames. These interfaces do not replace native detection, segmentation, or metric-depth evaluation and are excluded from unified LP/AP/FT rankings.

The comparison uses the same task prompts, vocabularies, and query locations for both frontier systems, with DINOv3 predictions mapped to the corresponding targets without retraining. Prompts prohibit external retrieval and request JSON predictions, validated by sample identifier before scoring. The verbatim prompts appear in `BASIC5_FRONTIER_SYSTEM_PROMPTS.md`.

For each dataset, the fixed 500-sample subset was divided into five non-overlapping batches of 100, each submitted in a separate prompt. Scores were computed over all 500 samples. Reproduction additionally requires the sample manifests, model identifiers and inference settings, raw outputs and retry records, and output-conversion and scoring procedures.