Distilling and Scaling Hierarchical Vision Transformer to 30B Parameters

Shuguang Dou[1] , **Dongqi Li**[2] , **Hao Jiang**[1] and **Dongsheng Jiang**[1*]
[1]Huawei Technologies Co., Ltd
[2]Institute of Computing Technology, Chinese Academy of Sciences
{doushuguang, jianghao66, jiangdongsheng1}@huawei.com, lidongqi24b@ict.ac.cn,

Abstract

Recent efforts have successfully scaled plain Vision Transformers (ViTs) to unprecedented sizes, ranging from 6B to 22B parameters. However, hierarchical ViTs, which are inherently better suited for multi-scale representation and dense prediction tasks, have largely remained constrained to under 2B parameters. This scaling gap prevents the community from leveraging high-capacity, multi-scale foundation models for complex visual scenes. To bridge this gap, we propose EHV, an Efficient Hierarchical ViT architecture designed for massive scaling. Our dense models scale from 200M to 5B parameters, and we further introduce a Sparse Mixture-of-Experts (SMoE) variant, pushing the scale to an industry-leading 30B parameters. The training pipeline follows a two-stage pretraining strategy: (1) MAE Stage: self-supervised pretraining on ImageNet-21K using a Masked Autoencoder (MAE); and (2) Distillation Stage: distilling knowledge from multiple state-of-the-art foundation models on a 27M-image dataset. With only 6.7B active parameters, EHV-5B-MoE achieves an exceptional 89.0% linear accuracy on ImageNet-1K, outperforming much larger models like EVA-CLIP-18B and DINOv3-7B. These results demonstrate that EHV effectively learns high-quality, generalizable, and linearly separable features across image classification, video analysis, and dense prediction tasks.

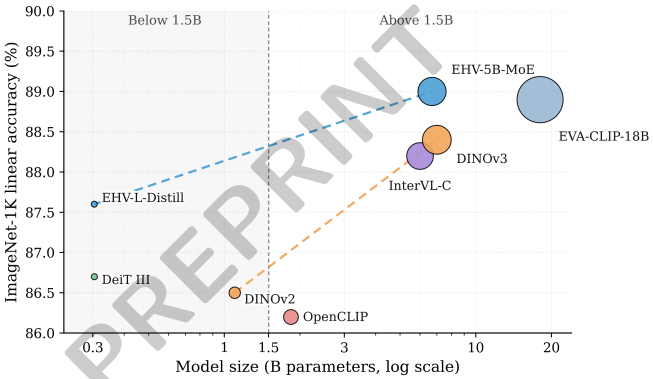

Figure 1: Scaling behavior of EHV with linear classification performance of forzen backbone on ImageNet-1K, compared with the current state-of-the-art and largest CLIP models and DINO series.

1 Introduction

Recent advancements in Vision Transformers (ViTs) have dramatically expanded model capacity, evolving from compact architectures to models with billions of parameters. Google's ViT-22B [Dehghani *et al.*, 2023] marked a milestone in end-to-end ViT scaling, achieving 22 billion parameters through optimized patch-based attention mechanisms and large-scale pretraining. InternVL [Chen *et al.*, 2024] subsequently demonstrated a balance between performance and efficiency at 6 billion parameters, while DINOv3-7B [Siméoni *et al.*, 2025] advanced self-supervised learning with a 7-billion-parameter architecture.

Despite these advances, research progress has largely concentrated on non-hierarchical plain Vision Transformers (ViTs). In contrast, to achieve stronger performance under fully supervised training on ImageNet-1K, many domain-specific visual Transformers adopt hierarchical designs, such as Swin [Liu *et al.*, 2021], MViT, and Hiera [Ryali *et al.*, 2023]. However, the largest existing hierarchical ViT, SwinV2-G [Liu *et al.*, 2022], contains only 3 billion parameters, which remains far below the scale of large language models (LLMs) and even plain ViTs.

To bridge this gap, we introduce EHV, the first efficient hierarchical ViT to reach 5 billion dense parameters and to further scale up to 30 billion parameters via Sparse Mixture-of-Experts (SMoE) techniques. Specifically, there are the following three points:

1) Architectural Optimization: To address the computational inefficiencies inherent in processing high-resolution features within hierarchical ViT, we introduce the Efficient Hierarchical Vision Transformer (EHV). Our optimization strategy is twofold. First, we integrate Linear ReLU Attention into the hierarchical architecture to reduce the quadratic complexity typically associated with self-attention mechanisms. Second, we fundamentally restructure the initial stage of the network by replacing the attention layers entirely with

*Corresponding Author

a Multi-Layer Perceptron (MLP). This design eliminates the excessive cost of attention operations at the early processing stage without compromising feature extraction capabilities. When coupled with Masked Autoencoder (MAE) pretraining, the proposed EHV architecture demonstrates superior efficacy. Empirical evaluations on downstream tasks, particularly object detection, indicate that EHV outperforms established baselines such as Hiera and Swin Transformer, validating the effectiveness of our efficiency-oriented architectural modifications.

2) Model Scaling: While Sparse Mixture-of-Experts (SMoE) models significantly expand model capacity by replacing standard Feed-Forward Networks (FFN) with expert layers, training such massive models from scratch is often computationally prohibitive and prone to instability. To mitigate these challenges, we adopt a progressive "upcycling" strategy that bridges dense pretraining with sparse finetuning. We begin by pretraining a stable 5-billion parameter dense model using the MAE framework to establish a robust representational foundation. Subsequently, we upgrade the model by converting its FFN layers into SMoE layers, each comprising 16 experts with a Top-2 gating mechanism. This transformation effectively scales the model to 6.67 billion activated parameters. By leveraging the converged weights of the dense model, this approach circumvents the cold-start instability issues typical of MoE training, resulting in a high-capacity model that balances training stability with inference efficiency.

3) Heterogeneous Knowledge Distillation: To further refine the capabilities of our EHV-Large and EHV-5B-MoE models, we propose a Heterogeneous Distillation Framework designed to synthesize diverse visual priors. Our framework leverages an ensemble of five state-of-the-art visual foundation models across distinct domains as teachers. We introduce specific distillation tokens and implement a mechanism to aggregate multi-scale features during the distillation process, facilitating a more effective knowledge transfer. This simple yet effective strategy yields significant performance gains. The resulting distilled models exhibit outstanding transfer learning capabilities, achieving state-of-the-art performance across a broad spectrum of vision tasks, including image classification, video classification, semantic segmentation, and object detection.

The main contributions are summarized as follows:

- We surpass the scaling limitations of hierarchical ViT by introducing EHV-5B-MoE, the largest model in the EHV family. The proposed architecture scales from 5 billion dense parameters to a total capacity of 30 billion parameters through the integration of a SMoE mechanism.
- We propose a robust training framework that combines self-supervised learning with multi-teacher knowledge distillation. By employing distillation tokens and multi-scale feature aggregation, EHV effectively absorbs knowledge from five diverse state-of-the-art visual foundation models, creating a highly versatile "student" representation.
- EHV demonstrates that massive scaling in hierarchical models leads to significant qualitative gains. With only 6.7B active parameters, it bridges the gap between classification-centric foundation models and downstream requirements, showing exceptional transferability across image/video classification, object detection, and semantic segmentation.

2 Related Work

2.1 Vision Foundation Model

The development of Vision Foundation Models (VFMs) originated with the success of deep Convolutional Neural Networks (CNNs) in the ImageNet challenge [He *et al.*, 2016]. The subsequent integration of the Transformer architecture [Dosovitskiy *et al.*, 2021] necessitated training on massive supervised datasets, such as the proprietary JFT-300M, which ultimately culminated in extremely large models like the ViT-22B encoder [Dehghani *et al.*, 2023].

Subsequent open-source efforts, such as OpenCLIP [Ilharco *et al.*, 2021], along with algorithmic refinements [Zhai *et al.*, 2023], have focused on improving data quality and loss design. Model scaling has been shown to be a critical factor; for example, EVA-CLIP-18B [Sun *et al.*, 2024] demonstrates that extending model capacity to 18 billion parameters yields highly robust visual representations and strong transfer learning performance. More recent works, including SigLIP 2 [Tschannen *et al.*, 2025] and the Perception Encoder [Bolya *et al.*, 2025], further confirm that performance is ultimately bounded by the availability of large-scale, high-quality data and substantial computational resources (e.g., training on more than 40 billion samples). InternVL-C [Chen *et al.*, 2024] exemplifies this trend by serving as a robust visual encoder that adopts a continuous learning paradigm to effectively bridge the representational gap between pure visual features and the latent space of language models, thereby providing a powerful visual backbone for modern multimodal architectures.

Concurrently, Self-Supervised Learning (SSL) has advanced significantly. DINOv2 [Oquab *et al.*, 2024] and DINOv3 [Siméoni *et al.*, 2025] series, through meticulous data curation, achieved performance parity with open-source weak-supervision methods. This strategy preserves the data efficiency intrinsic to SSL while enhancing the generality of the learned visual features.

2.2 Agglomerative Vision Foundation Models

The increasing specialization and diversity among modern VFMs have spurred a new area of research focused on consolidating disparate knowledge sources into unified, enhanced models—a paradigm we term Agglomerative Vision Foundation Models. This necessity arises from the desire to leverage complementary capabilities, such as the segmentation power of models like SAM and the zero-shot recognition of CLIP, within a single efficient framework.

Early efforts, exemplified by SAM-CLIP [Wang *et al.*, 2024a], directly addressed this challenge by integrating features from specific pre-trained models (SAM and CLIP) to combine their complementary strengths. Subsequent research has shifted toward more general and efficient knowl-

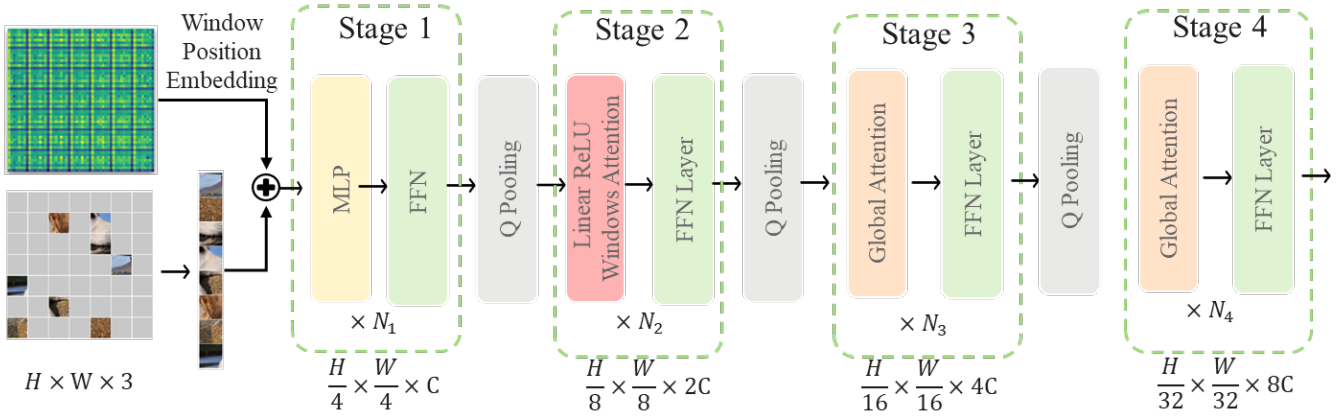

Figure 2: Hierarchical Stcrture of EHV. To improve efficiency, we employ an MLP and ReLU attention in the first and second stages, respectively, while applying a global attention mechanism in the remaining stages. At each stage transition, features from the Q layer and the skip-connection layer are doubled via a linear layer and then aggregated spatially using 2×2 max-pooling.

Setting	Acc@1↑
Hiera-L MAE	92.94
1. replace *att* in All Stages with *relu att*	91.88(-1.06)
2. replace *att* in Stages 1&2 with *relu att*	93.36(+0.38)
3. replace *att* in Stage 1&2 with *mlp*	92.26(-0.68)
4. replace *att* in Stage 1 with *mlp*	93.18(+0.22)
5. replace *abs pos* with *abs win pos*	93.28(+0.32)
6. combine 3 and 4	93.20(+0.26)
7. replace *att* in Stages 2 with *relu att*	93.56(+0.58)

Table 1: Linearing Hiera by replacing softmax attention with either MLPs or linear ReLU attention. Operation 7 is built upon Operation 6. All models are pre-trained using MAE for 100 epochs on ImageNet-100 and subsequently evaluated on ImageNet-100.

edge distillation paradigms. AM-RADIO [Ranzinger *et al.*, 2024] pioneered label-free knowledge distillation from multiple teacher models, enabling effective knowledge transfer without reliance on explicit supervision. UNIC [Sariyildiz *et al.*, 2024] introduced intermediate teacher-matching projectors along with a dynamic selection mechanism, allowing the student model to learn from the most relevant teacher at each stage of training. Theia [Shang *et al.*, 2024] further applied these multi-teacher distillation principles to downstream tasks, such as improving robotic learning by aggregating knowledge from multiple vision teachers.

Overall, agglomerative VFM research reflects a paradigm shift from training single, monolithic models toward constructing intelligent composite systems that synthesize specialized knowledge from a diverse ecosystem of foundation models.

3 Model Architecture

3.1 Efficient Hierarchical Vision Transformer

Our primary motivation was to introduce linear components to enhance the model's efficiency and transferability without compromising representational quality. We choose Hiera [Ryali *et al.*, 2023] as our base architecture, as its design without the Bells-and-Whistles. As shown in Table 1, we explored replacing the non-linear softmax attention with either a simple MLP or a linear ReLU attention:

- **Global Replacement**: Replacing *att* with *relu att* in *all* stages significantly degraded performance (91.88%, −1.06), suggesting that while non-linearity is detrimental in high-resolution stages, the full non-linear attention is crucial for deeper, low-resolution stages.
- **Targeted Replacement**: Focusing the linear replacement on the initial, high-resolution stages proved highly effective. Operation 7 (replacing *att* in Stage 2 with *relu att*, built upon Operation 6) achieved the highest accuracy of **93.56**% (+0.58). Operation 2 (replacing *att* in Stages 1 and 2 with *relu att*) also yielded a substantial improvement (**93.36**%, +0.38). This indicates that simplifying the self-attention mechanism in the early, computationally intensive stages effectively improves the feature quality. Replacing attention with a generic MLP in the early stages (Operations 3 and 4) showed mixed results. While replacing att in Stage 1 with mlp (Operation 4) resulted in a gain of +**0.22**%, replacing it in Stages 1 and 2 (Operation 3) led to a drop of −0.68%, suggesting that linear attention is superior to a simple MLP for replacing attention in the hierarchical context.
- **Positional Encoding:** We also investigated the role of positional encoding. Replacing the standard Absolute Position Embedding with Absolute Windowed Position Embedding in Operation 5 resulted in a performance gain (**93.28**%, +0.32). This suggests that incorporating local windowing constraints into the positional encoding further aids the model in capturing hierarchical and localized visual structures effectively.

These ablation studies confirm that the performance gains of EHV stem from a judicious application of linear mechanisms—specifically linear ReLU attention in the early high-resolution stages—combined with improved positional encoding schemes.

Model	Param	Blocks	Heads	Width	MoE
EHV-L	212M	2,6,36,4	2-4-8-16	144	✘
EHV-H	670M	2,6,36,4	4-8-16-32	256	✘
EHV-2B	2.07B	2,6,24,4	8-16-32-64	512	✘
EHV-5B	5.4B	2,6,28,4	8-16-32-64	768	✘
EHV-5B-MoE	30B	2,6,28,4	8-16-32-64	768	✔

Table 2: Configuration for EHV variants. Blocks and Heads specify number of blocks, and heads in each block for the four stages, respectively. Width denotes the channel width of first stage and the number of channels in each subsequent stage is twice that of the previous stage.

3.2 Scaling and Upcycling with Sparse MoE

The overall framework of EHV is illustrated in Fig. 2 and the scaling trajectory of the EHV series visual backbone is shown in Table 2. The design progresses from 0.2B parameters to 5B parameters, and ultimately achieves a leap to 30B parameters through the integration of an SMoE architecture, while preserving a consistent underlying structure.

Scaling Up. When scaling the model by increasing the number of blocks in the last two stages, we observed that excessively deep architectures (exceeding 70 layers) struggled to converge during MAE pretraining, in contrast to shorter and wider models. To address this, we adjusted the number of blocks in stage 3 (e.g., 36→24→28), while keeping the overall architectural configuration (2,6, X,4) unchanged. Channels and attention heads are systematically increased as the model scales, serving as the primary mechanism for enhancing model capacity. From EHV-L to EHV-5B, the parameter count increases nearly 25-fold, enabling the model to capture increasingly complex visual patterns.

Scaling Out with MoE. The EHV-5B-MoE model expands its parameter count from 5.4B to 30B (approximately a 5.5× increase) by replacing standard dense layers with Sparse Mixture-of-Experts (SMoE) layers, while preserving the depth, width, and head configurations of the EHV-5B backbone. Specifically, we employ a Top-2 routing mechanism where each input token x is dispatched to the two most relevant experts. The output y is computed as the weighted sum of these experts' outputs:

$$y = \sum_{i \in \text{Top-2}} G(x)_i E_i(x) \tag{1}$$

where $G(x)_i$ represents the gating probability assigned to expert E_i by the router, and $E_i(x)$ denotes the transformation applied by the i-th expert.

This design leverages the principle of sparse activation: although the total parameter count reaches 30B, only a subset of experts is activated for each input sample. Consequently, the model attains significantly greater expressive power and a higher capacity ceiling without incurring a proportional rise in computational cost, providing an efficient pathway toward ultra-large-scale hierarchical visual models.

4 Distillation with Vision Foundation Models

We employed five visual foundation models—DeiT III [Touvron *et al.*, 2022], dBOT [Liu *et al.*, 2024], AIMv2 [Fini *et al.*, 2025], PE-Core [Bolya *et al.*, 2025], and DINOv3 [Siméoni *et al.*, 2025]—as distilled teachers. To ensure efficient distillation, only large-scale models were selected as teachers. The EHV series includes models of various sizes; however, *due to limited computational resources, we only distilled two representative models: EHV-Large and EHV-5B-MoE.*

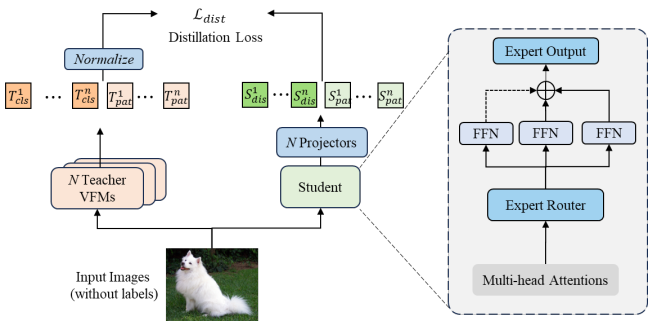

Figure 3: A simple yet effective multi-teacher heterogeneous distillation framework for training sparse MoE students.

4.1 Heterogeneous Ditillation Framework

Since EHV is a hierarchical ViT student model, while most visual foundation models adopt plain ViT architectures, the resulting heterogeneity poses challenges for distillation. Existing multi-teacher distillation methods typically assume homogeneous teachers, aligning features from corresponding layers in a one-to-one manner. However, such architectures are unsuitable for our setting. To address this, we introduce a simple and basic multi-teacher heterogeneous distillation framework, as illustrated in Fig. 3.

Each teacher is a vision foundation model (VFM) encoder $T = \{T^1, \cdots, T^n\}$ that maps an input image x into patch tokens T^i_{pat} along with a global CLS token T^i_{cls}. The objective is to learn the parameters of the student model S, such that its output representation matches the performance of the teacher models across all tasks. To achieve this, we employ a simple dimension alignment mechanism. For each teacher model, a distillation head h_{dill} is appended to the output of the student encoder, converting each token into a teacher-specific representation $h^i_{dill}(z)$.

Distillation loss. The distillation loss for each teacher is computed based on the output of its corresponding projection head. Similar to UNIC [Sariyildiz *et al.*, 2024], these distillation heads are auxiliary modules that facilitate training; they are removed after distillation and do not belong to the final student encoder. Each distillation head , h_{dist} is implemented as a Multi-Layer Perceptron (MLP), consisting of two linear layers and a GeLU nonlinearity.

To align student and teacher representations, we adopt a angular alignment loss. Angular alignment enforces directional consistency between the student and designated teacher features y. The angular alignment loss is formulated as:

$$\mathcal{L}_{dist} = -\frac{1}{P}\sum_{p=1}^{P} cos(y_p, h_{dist}(z)), \tag{2}$$

where $cos(\cdot, \cdot)$ denotes the cosine similarity and P denotes the patch count. This loss independently aligns the representations in both directions.

Load-Balancing Loss. To prevent the router from converging to a few "preferred" experts and ensure balanced utilization across the expert pool, we incorporate a load balancing loss $\mathcal{L}_{bal}$ [Fedus *et al.*, 2022]. For a batch of tokens, the loss is defined as:

$$\mathcal{L}_{lb} = \lambda \cdot M \sum_{i=1}^{M} f_i \cdot P_i \tag{3}$$

where M denotes the number of experts, f_i is the fraction of tokens dispatched to expert i, and P_i is the mean gating probability for expert i across the batch. The hyper-parameter λ controls the trade-off between the primary objective and load balance. By minimizing $\mathcal{L}_{lb}$, we encourage a uniform distribution of tokens (equal f_i) and a balanced confidence level (equal P_i) across all experts, thereby maximizing the model's total capacity and training efficiency.

Overall Loss. In conclusion, the total loss function for feature-aligned distillation to update the student model parameter is defined as:.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dist}} + \lambda \mathcal{L}_{\text{lb}} \tag{4}$$

4.2 Ditillation Tokens with Multi-scale Fusion

Due to the heterogeneity between teacher and student models, feature distillation cannot be performed at every layer as in isomorphic architectures. Aligning only the final-layer features prevents the student from effectively learning the teacher's rich multi-scale representations. Furthermore, since hierarchical ViTs lack CLS tokens, the mean of patch tokens is typically used to align with the teacher's CLS token during distillation. However, because the CLS token originates from global attention learning, such an alignment strategy may disrupt the joint learning of CLS and patch tokens.

Ditillation Tokens. To address the aforementioned issues, we introduce distillation tokens at last two stages of the multi-stage ViT architecture. These tokens are subsequently aggregated and aligned with the teacher's CLS token to facilitate more effective cross-level knowledge transfer. Specifically, the distilled tokens are concatenated with the patch tokens from each layer's input and subsequently interact through a self-attention mechanism. Since the feature dimensions vary across layers, the distilled tokens are not propagated to subsequent layers. Finally, the distilled tokens from all layers are projected to a unified dimension via pooling, summed, and aligned with the teacher-mapped features.

Classification with our approach. During inference, the distilled token functions analogously to a class token and is linked to a linear classifier to predict image labels. Unlike the multi-stage aggregation employed during the distillation phase, only the distilled token from the final layer is utilized for classification at inferene time.

Models	#Samples	Dataset
DINOv2	142M	LVD-142M
DINOv3	1689M	LVD-1689M
RADIOv2.5	1B	DataComp1B
EVA-CLIP-18B	2B	Merged-2B
ViT22B	4B	JFT
InterVL-C	6.03B	Multi open-source datasets
EHV-5B-MoE	27M	ImageNet-21K, CC3M, CC12M

Table 3: Details of training data for EHV-5B-MoE

5 Experiments

5.1 Training Details

Datasets. EHV-5B-MoE was distilled using only 27 million open-source images from datasets such as ImageNet-21K, CC3M, and CC12M. In contrast, EHV-Large relied solely on ImageNet-21K. As shown in Table 3, compared to previous visual foundation models, EHV required only a fraction of their training data volume, highlighting the remarkable efficiency of our distillation strategy.

Experimental setting. EHV-5B-MoE was trained using 196 visual tokens per image, where each token represents a 16×16 patch extracted from 224×224. EHV-5B-MoE is trained for 120 epochs with a batch size of 512 and BF16 precision. We adopt a cosine decay learning rate schedule with a peak value of 2×10^{-4} and apply linear warm-up during the first five epochs. Load-balancing loss is applied with $\lambda = 0.01$. The training process of EHV-5B-MoE was conducted using 32 GPUs, each equipped with 80 GB of VRAM, for a total duration of approximately 60 days.

5.2 Transfer to Image Classification

Efficient transfer learning with large-scale backbone networks is often achieved by employing them as frozen feature extractors. This section presents the evaluation results of EHV-5B-MoE on image classification tasks using linear probing and K-Nearest Neighbor (KNN).

Experimental Setup. For linear probing, we train on ImageNet using SGD with momentum for 40 epochs at a resolution of 224×224. Mild random cropping and horizontal flipping are applied as the only data augmentations, with no additional regularization. Similar to UNIC [Sariyildiz *et al.*, 2024], the learning rate and weight decay are determined through automatic parameter tuning.

Results on general image classification. EHV-L-Distill and EHV-5B-MoE exhibit exceptional performance on the ImageNet-1K linear evaluation benchmark, particularly within their respective parameter-scale categories. EHV-5B-MoE achieves a linear accuracy of 89.0%, the highest among all non-gray models in this group, outperforming EVA-CLIP-18B (88.9%) and DINOv3-7B (88.4%). It attains state-of-the-art results on both KNN (reflecting feature representation quality) and Linear (reflecting linear separability) evaluation metrics, demonstrating the model's ability to learn high-quality, generalizable feature representations.

Methods	Data	Model Size	ImageNet-1K *KNN*	ImageNet-1K *Linear*	ImageNet-V2
Below 1.5B parameter					
DeiT III [Touvron *et al.*, 2022]	IN21k	ViT-L	-	86.7	78.6
DINOv2 [Oquab *et al.*, 2024]	LVD-142M	ViT-g	83.5	86.5	78.4
RADIOv2.5 [Heinrich *et al.*, 2025]	DataComp1B	ViT-g	85.3	-	-
EHV-L-Distill (Ours)	IN21k	EHV-L	**86.7**	**87.6**	79.3
Above 1.5B parameter					
OpenCLIP [Ilharco *et al.*, 2021]	LAION-2B	ViT-G	83.2	86.2	77.2
InterVL-C [Chen *et al.*, 2024]	custom	6B	-	88.2	79.9
EVA-CLIP-18B [Sun *et al.*, 2024]	Merged-2B+	18B	-	88.9	79.3
DINOv3 [Siméoni *et al.*, 2025]	LVD-1689M	7B	-	88.4	81.4
EHV-5B-MoE (Ours)	Merged-27M	6.7B	**88.1**	**89.0**	**81.5**
ViT-22B [Dehghani *et al.*, 2023]	JFT-4B	22B	-	89.5	83.2
PE_{core}(448px) [Bolya *et al.*, 2025]	PEV-5B	2B	86.8	89.8	81.6

Table 4: Linear evaluation on ImageNet-1K with 224px resolutions. ViT-22B is trained on the private JFT-4B dataset, while PE_{core}-G uses a higher input resolution of 448 px. Therefore, both methods are marked in gray for fairness in comparison.

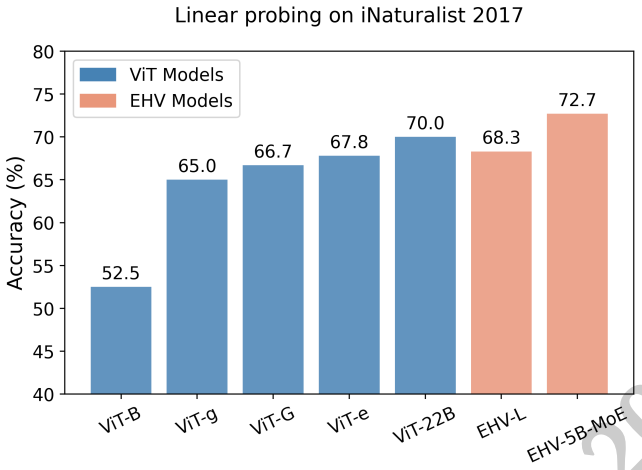

Figure 4: Linear probing on iNaturalist 2017 with 224px input resolutions. EHV-5B-MoE leads to significant accuracy improvement, especially when the input size is small.

Results on fine-grained classification. We further evaluated linear separability on the fine-grained classification dataset iNaturalist 2017 [Cui *et al.*, 2018], which contains 5,089 fine-grained categories distributed across 13 supercategories. Unlike ImageNet, this dataset exhibits a pronounced class imbalance, with a long-tailed distribution that presents greater challenges for classification. We compared EHV with flat Vision Transformers (ViTs), and the results are shown in Figure 4. The findings reveal substantial performance gains for EHV over flat ViTs—most notably, EHV-5B-MoE surpasses ViT-22B by 2.1%, despite the latter utilizing extensive private data and having roughly three times more active parameters.

5.3 Transfer to Video Classification

We assess the quality of the learned representations by transferring those obtained from pretraining the EHV-5B-MoE model on images to video classification tasks.

Experiment Setup. Following ViT-22B [Dehghani *et al.*, 2023], our video model first employs a spatial transformer that independently encodes each video frame. The distilled tokens from all frames are then input into a temporal transformer to capture temporal dependencies across frame representations. The spatial transformer is initialized with the pretrained weights from EHV-5B-MoE and subsequently frozen. This setup offers a computationally efficient approach for leveraging large-scale models in video tasks while effectively evaluating the representation quality learned by EHV-5B-MoE. Following ViT-22B, we sampled 128 frames from the Kinetics-400 dataset [Kay *et al.*, 2017] at a stride of 2 frames, and 32 frames from the Moments in Time dataset [Monfort *et al.*, 2020]. This setup differs from CoCa [Yu *et al.*, 2022], which assigns a label to each image patch in its video classification experiments and operates at a higher resolution of 576 px (compared to 224 px in our experiments), thereby producing substantially longer label sequences. We train 100 epochs with a batch size of 256 using AdamW with a learning rate of 1e-4 and with a cosine schedule including a linear warmup of 5 epochs.

Methods	Kinetetics 400	Moments in Time
RADINOv2.5	84.8	-
PE_{core}	87.9	-
DINOv3	87.8	-
CoCA	88.0	47.4
ViT-e	86.5	43.6
ViT-22B	88.0	44.9
EHV-5B-MoE	88.3	45.7
Fully finetuning SOTA	92.1	51.2

Table 5: Video classification results of frozen image representations. We evaluate the EHV-5B-MoE representations by freezing the backbone and training a small transformer to aggregate frozen, per-frame representations. Fully finetuning SOTA denotes InternVideo2 [Wang *et al.*, 2024b]

Results. As shown in Table 5, EHV-5B-MoE slightly outperformed ViT-22B and CoCA on K400. Although CoCA reported a marginally higher score of 47.4%, it was evaluated under different settings or with higher resolution. Specifically, EHV-5B-MoE is pre-trained exclusively on images (with a frozen backbone) and integrates frame-wise features using a lightweight sequential Transformer. Its ability to

Methods	ADE20K	Pascal VOC
PE_{core}	38.9	69.2
AM-RADIOv2.5	53.0	<u>85.4</u>
DINOv2-g	49.5	83.1
DINOv3-7B	**55.9**	**86.6**
ViT-22B	52.7	80.1
EHV-5B-MoE	<u>54.7</u>	84.3

Table 6: Dense linear probing results on semantic segmentation with frozen backbones. We report the mean Intersection-over-Union (mIoU) metric for the segmentation benchmarks ADE20k and VOC.

surpass ViT-22B on complex video classification tasks highlights the exceptional quality of its pre-trained image representations, which effectively capture and differentiate visual concepts while generalizing well to the sequential nature of video data. Despite its strong performance, EHV-5B-MoE still trails behind the fully fine-tuned SOTA model, InternVideo2 [Wang *et al.*, 2024b]. This performance gap is expected, as fully fine-tuned models such as InternVideo2 enable end-to-end optimization of the entire architecture on video datasets, thereby facilitating a more comprehensive adaptation to the spatio-temporal characteristics of video.

5.4 Transfer to Semantic Segmentation

Transfer learning for dense prediction is particularly important because obtaining pixel-level annotations is often expensive and labor-intensive. In this section, we evaluate the quality of the geometric and spatial information captured by the EHV-5B-MoE model through experiments on the semantic segmentation task.

Experiment Setup. We evaluate EHV-5B-MoE as a backbone for semantic segmentation on two benchmark datasets: ADE20K [Zhou *et al.*, 2017] and Pascal VOC [Everingham *et al.*, 2010]. A linear Transformer with 250 million parameters is trained on the frozen patch embeddings produced by EHV-5B-MoE, using a fixed input resolution of 512 pixels. All results are reported under a single-scale evaluation setting.

Results. As shown in Table 6, EHV-5B-MoE achieved an mIoU of 54.7 on ADE20K, ranking second overall (as underlined in the table). It substantially outperformed ViT-22B (52.7), DINOv2 (49.5), and PE_{core} (38.9). These results suggest that the hierarchical architecture and pre-training strategy (MAE combined with the MoE extension) of EHV-5B-MoE enable the model to learn high-quality visual representations that are highly effective for dense prediction tasks. On Pascal VOC, EHV-5B-MoE achieved an mIoU of 84.3, again surpassing ViT-22B's 80.1 and demonstrating robust performance. DINOv3 achieved the best performance on both datasets. This superiority can be attributed to its loss function design, which facilitates separable feature semantic scheduling and effectively exploits large-scale, high-quality training data.

5.5 Transfer to Object Detection

Object detection aims to predict bounding boxes for all object instances belonging to predefined categories. This task requires both accurate localization and robust recognition capabilities, as each bounding box must precisely align with object boundaries and correspond to the correct class. Although existing large-scale models have achieved over 66 mAP on standard benchmarks such as COCO, their substantial computational demands make 1B+ parameter visual models impractical for real-time applications. Therefore, we restrict our comparison to models that are fully fine-tuned at the large-parameter scale.

Bacbone	Pretrain	AP	AP_{50}	AP_{75}
Swin-L	Sup,ImageNet-21K	59.3	77.3	64.9
Hiera-L	MAE,ImageNet-1K	58.2	75.4	64.0
EHV-L	MAE,ImageNet-1K	59.8	77.0	65.6
EHV-L-D	Distill,ImageNet-21k	**61.2**	**79.1**	**67.5**

Table 7: COCO object detection using Co-DETR [Zong *et al.*, 2023]. All methods follow the training recipe from MMDetection.

Experiment Setup. We evaluated EHV-L's object detection capabilities on the COCO dataset [Lin *et al.*, 2014] and report results on the $COCO_{val}$ 2017 split. For implementation, we adopted the same training configuration as Swin-L [Liu *et al.*, 2021] in MMDetection, selected the 1× schedule, and fine-tuned it on COCO for 12 epochs.

Results. Despite being pre-trained solely with MAE self-supervised learning on the smaller ImageNet-1K dataset, EHV-L achieves 59.8 AP, slightly surpassing Swin-L (59.3 AP), which was pre-trained using large-scale supervised learning on ImageNet-21K. As an efficient hierarchical ViT model, EHV-L's MAE-pretrained features also substantially outperform those of Hiera-L, another self-supervised hierarchical ViT variant. EHV-L-D attains the best performance across all evaluation metrics in Table 7. This variant employs a distillation strategy and is pre-trained on the larger ImageNet-21K dataset, demonstrating that large-scale data and advanced pre-training techniques—such as distillation, which involves learning from a higher-performing teacher model—can further unlock the potential of the EHV-L architecture. While ViT-L achieves 65.9 mAP by pretrained on Object365, there remains considerable room for improvement for EHV-L using the same strategy.

6 Conclusion

In this work, we introduced EHV, a pioneering hierarchical vision transformer that breaks the long-standing scaling barrier of hierarchical architectures, reaching up to 30B parameters via Sparse MoE. By integrating architectural innovations with an efficient "upcycling" strategy, EHV maintains exceptional computational efficiency while maximizing model capacity. Furthermore, our heterogeneous distillation framework enables EHV to effectively consolidate knowledge from multiple foundation models. Experimental results demonstrate that EHV achieves state-of-the-art performance across a wide spectrum of tasks, including image and video classification, object detection, and semantic segmentation. This work establishes hierarchical ViTs as a powerful and scalable alternative for next-generation visual foundation models.

References

[Bolya *et al.*, 2025] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *CoRR*, abs/2504.13181, 2025.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024.

[Cui *et al.*, 2018] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge J. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, pages 4109–4118, 2018.

[Dehghani *et al.*, 2023] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 7480–7512, 2023.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georgios Melas, Albert Heigold, Sylvain Pham, Anselm Bertasius, Christoph Kelen, Jiang Lu, Neil Houlsby, and Michael Tschannen. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[Fedus *et al.*, 2022] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[Fini *et al.*, 2025] 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 M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *CVPR*, pages 9641–9654, 2025.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Heinrich *et al.*, 2025] Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2.5: Improved baselines for agglomerative vision foundation models. In *CVPR*, pages 22487–22497, 2025.

[Ilharco *et al.*, 2021] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, September 2021.

[Kay *et al.*, 2017] Will Kay, João 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. *CoRR*, abs/1705.06950, 2017.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, 2014.

[Liu *et al.*, 2021] 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 *ICCV*, pages 9992–10002, 2021.

[Liu *et al.*, 2022] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 11999–12009, 2022.

[Liu *et al.*, 2024] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. In *ICLR*, 2024.

[Monfort *et al.*, 2020] Mathew Monfort, Carl Vondrick, Aude Oliva, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa M. Brown, Quanfu Fan, and Dan Gutfreund. Moments in time dataset: One million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):502–508, 2020.

[Oquab *et al.*, 2024] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido 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. *Trans. Mach. Learn. Res.*, 2024, 2024.

[Ranzinger *et al.*, 2024] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: agglomerative vision foundation model reduce all domains into one. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 12490–12500, 2024.

[Ryali *et al.*, 2023] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 29441–29454, 2023.

[Sariyildiz *et al.*, 2024] Mert Bülent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. UNIC: universal classification models via multi-teacher distillation. In *ECCV (4)*, volume 15062 of *Lecture Notes in Computer Science*, pages 353–371. Springer, 2024.

[Shang *et al.*, 2024] Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 724–748, 2024.

[Siméoni *et al.*, 2025] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël 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, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025.

[Sun *et al.*, 2024] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. EVA-CLIP-18B: scaling CLIP to 18 billion parameters. *CoRR*, abs/2402.04252, 2024.

[Touvron *et al.*, 2022] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit III: revenge of the vit. In *ECCV (24)*, volume 13684 of *Lecture Notes in Computer Science*, pages 516–533, 2022.

[Tschannen *et al.*, 2025] Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025.

[Wang *et al.*, 2024a] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. SAM-CLIP: merging vision foundation models towards semantic and spatial understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshop*, pages 3635–3647. IEEE, 2024.

[Wang *et al.*, 2024b] 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 *ECCV (85)*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416, 2024.

[Yu *et al.*, 2022] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022, 2022.

[Zhai *et al.*, 2023] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 11941–11952, 2023.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, pages 5122–5130, 2017.

[Zong *et al.*, 2023] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *ICCV*, pages 6725–6735. IEEE, 2023.