VITRIX-UniViTAR: Unified Vision Transformer with Native Resolution

Limeng Qiao **Yiyang Gan** **Bairui Wang** **Jie Qin** **Shuang Xu** **Siqi Yang** **Lin Ma**✉

Meituan Inc.

qiaolm@pku.edu.cn, realgump@tju.edu.cn, {tjbairuiwang, jayqinliu}@gmail.com
sxu1997@126.com, siqi.yang@uq.net.au, forest.linma@gmail.com

Abstract

Conventional Vision Transformer streamlines visual modeling by employing a uniform input resolution, which underestimates the inherent variability of natural visual data and incurs a cost in spatial-contextual fidelity. While preliminary explorations have superficially investigated native resolution modeling, existing works still lack systematic training recipe from the visual representation perspective. To bridge this gap, we introduce **Uni**fied **Vi**sion **T**ransformer with N**A**tive **R**esolution, *i.e.* UniViTAR, a family of homogeneous vision foundation models tailored for unified visual modality and native resolution scenario in the era of multimodal. Our framework first conducts architectural upgrades to the vanilla paradigm by integrating multiple advanced components. Building upon these improvements, a progressive training paradigm is introduced, which strategically combines two core mechanisms: *(1)* resolution curriculum learning, transitioning from fixed-resolution pretraining to native resolution tuning, thereby leveraging ViT's inherent adaptability to variable-length sequences, and *(2)* visual modality adaptation via inter-batch image-video switching, which balances computational efficiency with enhanced temporal reasoning. In parallel, a hybrid training framework further synergizes sigmoid-based contrastive loss with feature distillation from a frozen teacher model, thereby accelerating early-stage convergence. Finally, trained exclusively on public accessible image-caption data, our UniViTAR family across multiple model scales from 0.3B to 1.4B achieves state-of-the-art performance on a wide variety of visual-related tasks. The code and models are available here.

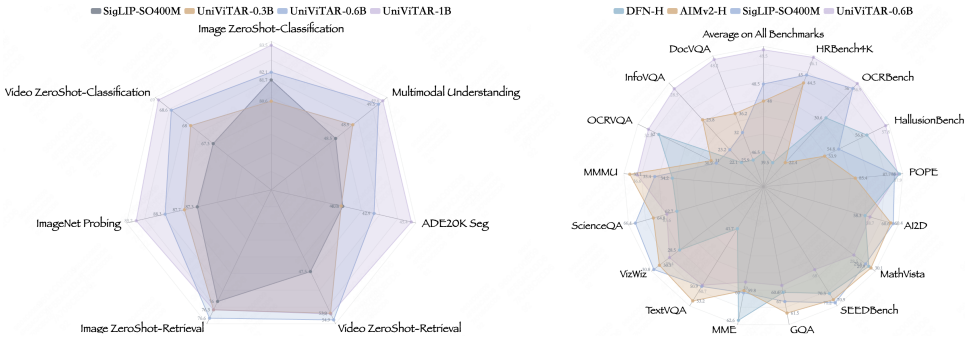


Figure 1: The figure presents: (left) a systematic overview of model scaling performance across downstream tasks when increasing parameter size from 0.3B to 1B, and (right) a comprehensive comparison of multimodal capabilities against SOTA baselines on diversified benchmarks.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

In the era of rapid advancements of multimodal large models, Vision Transformer [1], characterized by its simplicity and scalability, has emerged as a foundational architecture for visual representation learning. Drawing inspiration from transformer-based large language models, conventional ViT usually uniformly converts raw visual data into square aspect ratio and standardized resolution to reduce modeling complexity and simplify the training workflow. While this paradigm simplifies feature extraction and aligns with existing engineering practices, it inherently imposes artificial constraints on real-world visual data by disregarding the inherent variability of natural images.

Recent studies have preliminarily investigated the vision backbone within a native resolution paradigm. FlexViT [2] introduces a flexible ViT architecture featuring dynamical patch size selection in the patch embedding layer, which facilitates smooth variation of token sequence length through parametric scaling. In contrast, NaViT [3] maintains fixed patch size while directly processing native resolution images with varying aspect ratios, where the token sequence length of different images changes dynamically. This approach demonstrates the feasibility and benefits of adopting natural language processing style packing strategies for vision foundational model. Qwen-VL's [4, 5] vision encoder inherits NaViT's core configuration while specifically investigating native resolution impacts from a multimodal large model perspective. While the aforementioned approaches have attracted initial research attention, the field still lacks a comprehensive series of architecture-homologous vision backbones that can simultaneously support native- and fixed-resolution processing, achieve high-fidelity feature extraction for both images and videos.

To address this gap, we present the **Uni**fied **Vi**sion **T**ransformer with N**A**tive **R**esolution, termed as **UniViTAR**, a family of vision foundational backbones designed to uniformly process visual modalities (image or video) with native resolution and dynamic aspect ratio. Building upon insights from large language model recent practices and architectural innovations in visual transformers, our approach firstly conduct systematic architectural upgrades to the vanilla ViT paradigm by integrating multiple advanced components: 2D Rotary Position Embedding, SwiGLU activation function, RMSNorm layer, QK-Norm mechanism, and LayerScale module. These modifications collectively establish a more robust architectural foundation compared to conventional implementations. Secondly, we develop a progressive training paradigm with two complementary adaptation strategies: *1)* the progressive resolution adaptation strategy employs curriculum learning from fixed low-resolution (e.g., 224) pretraining to native-resolution fine-tuning. Notably, our experiments reveal that the advanced ViT architecture exhibit remarkable adaptability - models pretrained at fixed resolution can efficiently generalize to variable-length visual sequences through limited native resolution tuning. *2)* the progressive visual modality adaptation strategy addresses computational challenges in video processing by deferring video data integration to the final training phase. We further demonstrate that alternating image-video training sequences (inter-batch modality switching) significantly outperforms mixed-batch (intra-batch modality mixing) in preserving image understanding capabilities while acquiring temporal reasoning skills. Thirdly, we implement a hybrid training framework combining contrastive learning objectives with distillation techniques. Our primary optimization employs a sigmoid-based contrastive loss [6] for unified image-video representation learning. To accelerate early-stage convergence, we further incorporate feature distillation from a frozen vision teacher model as an auxiliary training objective during initial phases, then gradually phasing out this regularization as the model matures. Finally, through this comprehensive approach trained on public-accessible datasets, we successfully scale a family of vision backbones supporting native resolutions and both visual modalities, with parameter counts ranging from 0.3B to 1.4B. Extensive evaluations demonstrate the effectiveness of our proposed methods.

Specifically, the contributions of our UniViTAR family are summarized as follows:

- We introduce a family of homogeneous visual foundation models that support native resolution and unified feature extraction across visual modalities, offering the community a versatile framework for multimodal research.
- We develop an efficient and effective progressive training strategy that addresses the computational challenges of native resolution modeling while systematically enhancing the model's image-caption alignment capability.
- We train our models with public-accessible datasets, achieve leading performance with limited resources, and observe a trend of performance increasing with parameter scaling.

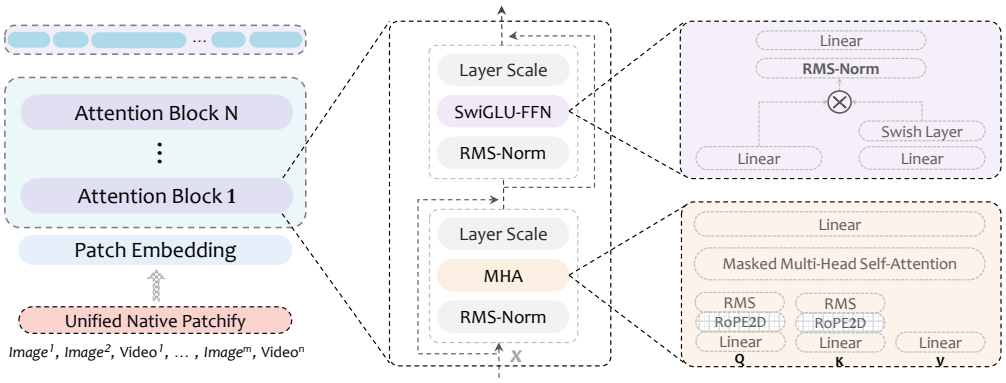


Figure 2: **Architecture of UniViTAR family.** All visual inputs are uniformly transformed into patch sequences and fed into Vision Transformer. In addition to using the *Pre-Norm* approach, we also adopt *RMS-Norm* as the normalization layer in both *MHA* and *FFN* module.

2 Method

2.1 UniViTAR: Homologous Visual Foundation Model

2.1.1 Architecture Design

UniViTAR is a Transformer-based encoder model that inherits the original architecture of the conventional Vision Transformer [1] but incorporates the following advanced modifications:

Unified Patchify for Native Image and Video Modality. As illustrated in the Figure 2, given the native input visual data $\mathbf{X} \in \mathbb{R}^{T \times C \times H \times W}$ of any vision modality (image, video), where $T = 1$ represents image and $T > 1$ represents video, **UniViTAR** firstly patchifies $\mathbf{X}$ into a series of dynamic length visual patch sequences $\mathbf{P} = (N, S)$, where N is the number of patches per image/video and S is the number of pixels per patch. Then a $3D$ convolution layer is adopted as the *Patch Embedding Layer* to consistently convert the above patch sequence into a visual token sequence $\mathbf{T} = (N, D)$, where D is the hidden size of the following attention layers.

2D Rotary Position Embedding. Drawing on the architecture designs of language models, the original ViT regards the position information among different visual tokens as a one-dimensional association. In fact, considering that visual data usually has spatial association (row and column) and temporal association (time), the position information between different tokens is usually considered to be multi-dimensional. Thence we remove the original absolute position encoding and introduce $2D$-RoPE [7] into each subsequent encoder layer to capture the two-dimensional positional information of images. Furthermore, we found that the presence or absence of the class token in the original ViT has almost no effect on model performance. To ensure the consistency of position encoding, we also empirically remove the design of class token.

SwiGLU and RMSNorm. By leveraging the recent advances of LLaMA [8] architecture design for language modeling, UniViTAR incorporates SwiGLU as the feed-forward network (FFN) and replaces all normalization layers with RMSNorm. In addition, we adds an extra RMSNorm to each SwiGLU-FFN for good expressivity and improving the training stability.

Query-Key Normalization. In order to improve the stability of model training, we adopt the QK-Norm technique [9, 10] , which applies normalization to the queries and keys before the dot-product attention computation, to directly controls the norm growth of input to the softmax and avoid abnormal attention logits. Note that we still utilize RMSNorm as the norm function to ensure the consistency of the architecture.

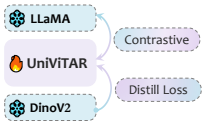
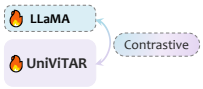
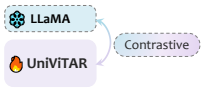
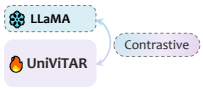
2.1.2 Homologous Model Scaling

The UniViTAR family consists of a comprehensive suite of foundational and scratch-train models, encompassing a parameter range from 0.3 to 1.4 billion, *i.e.* UniViTAR-0.3B/0.6B/1B. The hyper-parameters and important information are listed in Table 1 in details.

Table 1: **Detailed architectural configuration for UniViTAR family.**

Model	Hidden-Size	Intermediate-Size	Encoder-Layers	Attention-Heads	Parameters (M)
UniViTAR-0.3B	1024	4224	24	16	**310**
UniViTAR-0.6B	1280	5184	32	16	**637**
UniViTAR-1B	1920	7680	32	24	**1419**

Table 2: **Detailed training strategy illustration of UniViTAR family.**

	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**
Train Strategy				
Data Modality	Image	Image	Image	Image, **Video**
Resolution	224×224	224×224	**Native**	**Native**
Loss Function	Sigmoid, KL	Sigmoid	Sigmoid	Sigmoid
Seen Samples	12B	1B	1B	0.6B

2.2 Contrastive Vision-Language Pretrain with UniViTAR

2.2.1 Architecture Design

In general, the acquisition of UniViTAR largely follows the basic training paradigm of CLIP [11]. Specifically, the native-resolution visual input v is encoded into the visual feature space via the UniViTAR encoder to obtain $F_v \in \mathbb{R}^{N_v \times D_v}$, while the textual input t is projected into the textual feature space through a pretrained LLaMA [12] decoder to obtain $F_t \in \mathbb{R}^{N_t \times D_t}$. The dynamic-length visual features F_v are then uniformly converted into the visual embedding $f_v \in \mathbb{R}^{D_v}$ through a global average pooling and the feature corresponding to the <*EOS*> token in F_t is utilized as the textual representation $f_t \in \mathbb{R}^{D_t}$ of the input caption. Subsequently, f_v and f_t are further projected into the same shared semantic space via a linear projection layer respectively. Then a simple pairwise sigmoid loss [6] is employed as the contrastive supervision to align the visual and text modalities semantically.

2.2.2 Optimized Contrastive Training Strategy

To ensure that the model can converge efficiently and the training cost is controllable, we carefully design the training pipeline of UniViTAR into four stages in sequence, as shown in Table 2.

Stage 1: Visual knowledge pre-acquisition with hybrid paradigm training. The primary objective of this phase is to efficiently pretrain a visual foundation model from scratch by integrating two classic learning paradigms: vision-text contrastive learning and visual knowledge distillation. Specifically, the proposed architecture employs a triple-branch parallel design: (1) a *xavier*-initialized UniViTAR, (2) a frozen pre-trained text encoder, and (3) a frozen pre-trained visual teacher. During training, only the target visual foundation model receives gradient updates, with other branches fixed to minimize computational overhead while preserving knowledge integration. For implementation, we adopt LLaMA [12] and DINOv2-g [13] as default components, though the framework supports substitution with alternative pre-trained foundation models. The composite training objective is defined as:

$$\mathcal{L}_{overall} = \mathcal{L}_{contrastive}(f_v^{UniViTAR}, f_t^{LLaMA}) + \lambda \cdot \mathcal{L}_{distillation}(f_v^{UniViTAR}, f_v^{Dino}) \quad (1)$$

where $\mathcal{L}_{contrastive}$ is the sigmoid loss from SigLIP [6] and $\mathcal{L}_{distillation}$ is the KL Divergence [14]. The target visual foundation model functions as a visual knowledge bridge, simultaneously performing image-text alignment and feature distillation. This phase processes 12B samples with all images resized to 224, constituting 82.2% of the total training data (12B/14.6B).

Stage 2: Finetune with full-parameter for superior alignment. The objective of this stage is to further enhance the upper limit of image-text alignment through full-parameter fine-tuning of both vision and text encoders, establishing a unified semantic-visual space. The visual distillation branch is deactivated during optimization. Training employs identical image-caption pairs as Stage 1 at 224 resolution. Considering the high computational cost of full parameter fine-tuning, the training process is conducted on 1B samples, accounting for 6.9% of the total training data.

Stage 3: Unlock the model-capacity of native-resolution. In this stage, our strategy extends the model capability to handle native-resolution, thereby achieving robust image-text alignment for dynamic-resolution inputs. However, enabling native-resolution capacity necessitates addressing two critical challenges: (1) ensuring positional encoding are thoroughly trained across variable sequence lengths, and (2) transfering feature distribution from uniformly-resized patches to native patches through efficient training. In practice, visual data is batched in its native form to preserve original resolutions and aspect ratios. Then the intra-batch images are dynamically scaled (with aspect ratios maintained) to align total sequence lengths L_{total} with a predefined token limit L_{max}. That is to say, when the value of L_{max}/L_{total} is greater than 1, all data will be uniformly enlarged, and vice versa the shape of all data will be reduced, ensuring consistent computational loads across batches. Within attention blocks, each token's receptive field is confined to tokens from the same image via masking, enabling isolated intra-image contextual modeling while preserving inter-sample independence. During training, resolution diversity within batches ensures comprehensive training of positional encoding across varying context lengths, progressively refining the model's ability to generate features aligned with native patch distributions. At inference, inputs are processed directly at their native resolutions without resizing. In this stage, 1B samples (6.85% of the total training) were trained with the text branch frozen throughout the process.

Stage 4: Unifying visual modalities with image-video alternation training. The goal of this stage is to unify image and video input modalities with native-resolution and dynamic video length. Inspired by the InternVideo series [15, 16], we utilize both image-text and video-text pairs to optimize the UniViTAR checkpoint from Stage 3 with an image-video alternating training strategy. This strategy addresses three critical considerations: (1) leveraging image data's superior scale and diversity compensates for video data scarcity while maintaining visual content continuity; (2) joint image-video training preserves cross-modal comprehension capabilities; (3) alternating modality-specific updates enforce focused parameter adaptation through sequential modality optimization. The alternating training protocol first initiates each epoch with random permutation of image-video data to enhance stochasticity. Subsequently, data batches are partitioned into global batch units and alternately sequenced at global batch granularity. Through this structured approach, the configuration effectively maintains modality purity within individual training batches by enforcing strict image-video alternation. To accommodate native-resolution video processing with dynamic lengths, we implement adaptive frame sampling: full temporal retention when frame count $F < F_{max}$, and uniform subsampling to F_{max} frames when exceeded. With the predefined token constraints (L_{min}, L_{max}) and the calculated frame length F, all frames are subsequently resized within these computed bounds while preserving original aspect ratios.

2.3 UniViTAR as a Vision Encoder for MLLMs.

In this section, we introduce a simple strategy for constructing an effective native resolution MLLM based on the UniViTAR series. The common and industry-validated Vision-Language Models (VLMs) paradigm typically combines pretrained visual backbones with large language models, followed by multimodal training on a rich mixture of vision language tasks. To ensure fair comparison and minimize bias, we adhere to this established configuration. Specifically, we employ UniViTAR as the vision encoder and employ Qwen2.5-1.5B [5] as the large language model. Following established practices [17], we implement a three-layer *MLP* with pre-normalization and a $2\times$ pixel-unshuffle operation [18] along the width dimension as the vision-language adapter to bridge the visual and linguistic modalities. For native-resolution modeling, we identify two primary challenges. On one hand, due to the varying lengths of input samples, the boundary between vision and language tokens is not fixed. To enhance "modality isolation", we introduce specialized prompts, known as *Boundary Markers*, such as <*image_start*> and <*image_end*>, at the beginning and end of the vision token sequence. On the other hand, 2D-to-1D flattening of vision tokens may compromise the information of the height-width ratio. To mitigate this, we incorporate *Line Anchors*, such as <line-*idx*>, into the vision tokens, where *idx* denotes the corresponding vertical positions in the original patchified image, thereby potentially strengthening positional awareness in compressed tokens. For a vision token sequence of length hw, the original arrangement $x^{1,1}, ..., x^{1,w}, ..., x^{2,w}, ..., x^{h,w}$ is transformed as:

$$<image\_start>, x^{1,1}, \ldots, x^{1,w}, <line{-}1>, x^{2,1}, \ldots, x^{h,w}, <line{-}h>, <image\_end> \quad (2)$$

Notably, these added markers are string-based identifiers rather than special tokens of the tokenizer. To systematically evaluate multimodal comprehension capabilities, we adopt a dual-stage training paradigm motivated by established methodologies in vision-language alignment like [19, 20].

3 Experiments

3.1 Training Recipe

Data Details. We collect public accessible image-text pairs and build our Merged-1B dataset, which is composed of DataComp-1B [21], COYO [22], LAION-2B [23], LAION-400M [24], DFN-2B [22], CC12M [25] and CC3M [26]. Moreover, to further enhance the video feature extraction capabilities of UniViTAR, we meticulously constructed a dataset Merged-65M of roughly 65 million samples by randomly selecting video clips from three public accessible video datasets, *i.e.*, Panda-70M [27], WebVid-10M [28], and InternVid-10M-FLT [29]. We refer to the combined image and video data mentioned above as Merged-1.1B. The detailed data composition is summarized in the Appendix.

Hyperparameter Details. The detailed hyperparameter configurations for each training stage are presented in the Appendix. As tabulated, we utilize a progressive reduction of the peak learning rate in correlation with increasing visual backbone scale to ensure optimal training stability. Notably, the learning rate of text branch in Stage 2 remains consistently one-tenth of the visual component throughout this phase. To enhance training efficiency, we integrated the DeepSpeed library [30] by employing ZeRO optimizer sharding [31], gradient checkpointing [32], and flash attention [33].

3.2 Results on Zero-shot Image Classification & Retrieval

Evaluation Setup. Our evaluation protocol encompasses both zero-shot classification and cross-modal retrieval tasks. For zero-shot classification, we conduct evaluation on ImageNet [34] and its established variants [35, 36, 37, 38, 39]. Each class is represented by multiple text prompts curated from [11, 40]. The *Top-1* accuracy is utilized to evaluate the model performance. For cross-modal retrieval assessment, we adopt the benchmark protocols defined in [41], evaluating on Flickr [42] and MS-COCO [43] using their official partitions. The retrieval paradigm involves bidirectional image-text matching, namely image-to-text retrieval and text-to-image retrieval tasks.

Results Comparison and Analysis. Table 3 demonstrates the exceptional performance of our model at comparable parameter scales. As the model size increases from 0.3B to 1.4B, the average zero-shot classification accuracy across six benchmarks exhibits a progressive improvement trend, rising from 80.5% to 81.9% and further to 83.4%. Notably, all models of varying scales employ identical training samples and strategies, with this performance enhancement attributed to *parameter scaling* effects—a finding consistent with established scaling laws in transformer-related research. As detailed in the table, our UniViTAR-1B shows superior performance despite utilizing a smaller training corpus, outperforming its counterparts with more parameters, such as InternViT-6B [12] and EVA-8B [44]. We posit that this advantage stems from two key factors: optimized model atchitecture and training strategy, and preservation of native input resolution, which generates higher-quality visual tokens.

3.3 Results on Zero-shot Video Classification & Retrieval

Evaluation Setup. We evaluate the zero-shot video classification performance on three popular benchmarks as K-400 [50], UCF-101 [51] and HMDB51 [52], using the class names as text prompts. Also, we evaluate the zero-shot video-text retrieval performance on ActivityNet [53], MSR-VTT [54] and MSVD [55]. Following [15, 16], for each video in the 1K version of the test split, we sample one sentence from every set of 20 sentences for MSR-VTT. Following [56], we concatenate the multiple descriptions to form a paragraph and perform a paragraph-to-video retrieval on ActivityNet. All videos are sampled with a dynamic frame rate, with each frame dynamically resized to maintain the original aspect ratio while ensuring the total token within the range of 576 to 16,384.

Results Comparison and Analysis. Table 4 shows the performance of our UniViTAR series models on video benchmarks across comparable parameter scales. As the model size scales from 0.3B to 1B, UniViTAR exhibits consistent performance gains on video benchmarks, with average zero-shot classification metrics improving from 68.0 to 69.0. When compared to models trained on image-caption data under similar parameter scales, UniViTAR achieves notable improvements. These advancements can be attributed to two key design choices: (1) preserving the aspect ratio of each frame to retain the original semantic information of visual content, and (2) employing dynamic video frame sampling to effectively capture detailed temporal information. However, when compared to the models trained exclusively on video-caption data, UniViTAR still has room for improvement compared to some of the latest models [57, 58, 16], as shown in the Table 4 with gray color.

Table 3: **Evaluation of zero-shot performance on various image benchmarks**. The symbol 👁‍🗨 indicates that the image-caption data used by the corresponding method is not publicly available.

Method	Data Source	Res.	Overall	ImageNet Variants						Overall	Flickr		COCO	
				IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net		T→I	I→T	T→I	I→T
CLIP-L [11]	WIT400M 👁‍🗨	224	72.1	75.5	70.8	87.8	69.8	59.6	68.9	60.8	65.0	85.2	36.5	56.3
OpenCLIP-L [45]	DataComp1B	224	75.7	79.2	69.6	90.8	72.1	68.0	74.3	67.9	73.4	89.0	45.7	63.3
MetaCLIP-L [46]	CC-2.5B 👁‍🗨	224	76.6	79.2	72.3	92.1	72.6	69.0	74.6	69.5	76.4	90.1	47.1	64.4
DFN-L [47]	DFN5B 👁‍🗨	224	77.1	**82.2**	67.5	91.8	75.7	70.4	74.8	69.8	75.5	89.6	48.6	65.6
EVA02-L [44]	Merged-2B	336	77.5	79.8	76.2	92.7	73.0	68.1	74.9	69.9	78.0	89.6	47.9	64.2
CLIPAv2-L [48]	DataComp1B	336	78.1	80.3	77.7	93.3	73.5	70.9	73.1	69.5	74.6	90.4	47.2	65.6
SigLIP-L [6]	WebLI10B-En 👁‍🗨	384	79.4	82.1	76.6	**95.1**	**75.9**	**73.6**	72.8	75.2	81.4	93.7	53.9	**71.9**
UniViTAR-0.3B	Merged-1B	Native	**80.6**	81.5	**84.1**	93.9	75.1	69.7	**79.1**	**76.3**	**84.0**	**95.1**	**54.7**	71.2
OpenCLIP-H [45]	LAION2B-en	224	72.3	78.0	59.4	89.3	70.9	66.6	69.4	68.7	75.5	89.5	46.5	63.4
MetaCLIP-H [46]	CC-2.5B 👁‍🗨	224	78.4	80.5	75.3	93.4	74.2	70.5	76.4	71.3	78.3	91.8	48.8	66.2
CLIPAv2-H [48]	DataComp1B	336	80.8	81.8	82.7	94.4	75.6	72.8	77.4	70.8	76.3	90.3	49.2	67.2
DFN-H [47]	DFN5B 👁‍🗨	378	80.5	**84.4**	79.6	93.8	**78.3**	73.2	73.4	75.9	82.0	94.0	**55.6**	71.9
SigLIP-SO [6]	WebLI10B-En 👁‍🗨	384	81.7	83.1	82.5	**95.8**	77.2	**74.5**	77.0	76.0	83.0	94.3	54.2	**72.4**
UniViTAR-0.6B	Merged-1B	Native	**82.1**	82.3	**86.8**	94.9	76.1	71.6	**81.1**	**76.6**	**84.1**	**95.5**	55.4	71.7
OpenCLIP-g [45]	LAION2B-en	224	73.0	78.5	60.9	90.2	71.6	67.5	69.1	71.1	77.7	91.4	48.8	66.4
OpenCLIP-G [45]	LAION2B-en	224	76.2	80.1	69.3	92.1	73.6	68.9	72.8	72.8	79.6	92.9	51.4	67.4
EVA01-g [49]	Merged-2B	224	76.9	79.3	74.2	92.5	72.1	68.1	74.9	72.3	79.0	91.7	50.3	68.2
EVA02-E [44]	Merged-2B	336	80.9	82.0	82.2	94.6	75.6	71.6	79.4	73.2	78.9	94.1	51.1	68.7
CLIPAv2-G [48]	DataComp1B	336	82.7	83.1	86.0	95.4	77.3	**74.5**	79.7	72.2	78.3	92.2	50.4	67.8
InternViT-6B [12]	InternVL-5B	224	82.5	83.2	83.8	**95.7**	77.3	74.3	80.6	75.3	81.7	94.7	54.1	70.6
EVA-8B [49]	Merged-2B	224	82.9	**83.5**	85.2	95.3	**77.7**	74.3	81.2	74.9	80.8	**95.6**	53.0	70.3
UniViTAR-1B	Merged-1B	Native	**83.5**	82.9	**89.1**	**95.7**	77.3	73.4	**82.8**	**76.3**	**83.5**	95.1	**55.3**	**71.3**

Table 4: **Evaluation of zero-shot performance on various video benchmarks**. The symbol † signifies that the reported metrics are based on our own evaluations.

Method	Type	Res.	Frames	Overall	Classification			Overall	ANet		MSR-VTT		MSVD	
					K400	UCF	HMDB		V→T	T→V	V→T	T→V	V→T	T→V
†OpenCLIP-L [45]	Image	224	16	58.4	61.5	69.2	44.5	41.0	32.0	34.2	30.1	37.5	63.7	48.5
†DFN-L [47]	Image	224	16	56.4	56.8	67.7	44.8	40.4	31.6	34.1	32.1	35.2	61.9	47.7
†EVA02-L [44]	Image	336	16	64.4	64.4	76.0	52.8	44.7	35.8	37.2	35.4	39.7	69.1	51.0
†SigLIP-L [6]	Image	384	16	64.8	64.2	79.2	50.9	45.3	34.3	35.8	35.7	40.0	73.0	**53.0**
ViCLIP-L [29]	Video	224	8	-	64.8	-	-	41.2	24.0	15.1	41.3	42.4	75.1	49.1
InterVideo-L [15]	Video	224	16	-	64.3	80.5	-	42.2	31.4	30.7	39.6	40.7	67.5	43.4
UMT-L [59]	Video	224	16	-	-	-	-	47.7	39.4	41.9	38.6	42.6	74.5	49.0
UniViTAR-0.3B	Image&Video	Native	2∼32	**68.0**	**66.0**	**82.6**	**55.4**	**53.9**	**47.9**	**49.9**	**48.0**	**48.8**	**77.8**	50.7
†OpenCLIP-H [45]	Image	224	16	62.0	61.7	72.5	51.6	43.5	36.1	38.9	34.5	38.9	63.3	49.4
†DFN-H [47]	Image	378	16	62.9	63.8	76.7	48.2	46.2	39.7	42.9	36.1	39.6	66.6	52.4
†SigLIP-SO [6]	Image	384	16	67.3	66.8	**83.0**	52.1	47.5	36.6	39.3	37.5	41.1	75.5	**54.7**
TVTSV2-H [60]	Video	224	12	63.2	59.6	78.0	52.1	-	-	-	-	41.3	-	-
UniViTAR-0.6B	Image&Video	Native	2∼32	**68.6**	**67.6**	82.9	**55.2**	**54.9**	**48.7**	**51.5**	**48.6**	**50.2**	**75.8**	54.3
†OpenCLIP-g [45]	Image	224	16	63.1	61.5	76.6	51.1	44.4	36.8	39.8	36.4	39.2	64.3	50.1
†OpenCLIP-G [45]	Image	224	16	64.2	63.2	76.2	53.4	46.0	36.7	41.4	36.9	41.8	67.5	51.5
†EVA01-g [49]	Image	224	16	62.8	63.4	72.1	52.9	45.5	37.0	40.1	37.2	40.1	67.6	50.8
InternViT-6B [12]	Image	224	8	-	**69.1**	-	-	-	-	-	42.4	46.3	-	-
UniViTAR-1B	Image&Video	Native	2∼32	**69.0**	68.6	**81.0**	**57.3**	54.0	47.8	49.6	48.3	47.6	75.5	55.2
VideoCoCa-g [57]	Video	224	8	72.4	72.0	86.6	58.7	39.0	33.0	34.5	64.7	34.4	33.0	34.5
VideoPrism-g [58]	Video	288	16	-	76.4	-	-	-	50.3	52.7	51.7	52.7	-	-
InternVideo2-6B [16]	Video	224	8	-	-	-	-	62.0	56.5	63.2	53.7	55.9	83.1	59.3

3.4 Results on Image Classification by Linear Probing

Following common prectices [12, 61], we assess the performance of UniViTAR family as off-the-shelf backbones on image classifications. Specifically, we train a linear classifier on the last feature layer with a frozen backbone on ImageNet-1K [34] and evaluate the performance on the validation set and other ImageNet variants [62, 35, 36, 37, 38]. In addition, we also report the classification performance with attentive probing setting as used in [61], which adopts a cross-attention layer with random initialized queries. Table 5 represents the downstream classification performance of our models. First, as the model size increases, the average performance across six benchmarks demonstrates consistent improvement. Second, we observe that the attentive probing performance shows stable improvements over linear probing. Furthermore, compared to public methods, our UniViTAR family shows superior performance across various parameter scales.

Table 5: **Evaluation of classification performance on various image benchmarks**. The † signifies that the reported metrics are based on our own evaluations.

Method	Classifier	Res.	Overall	ImageNet Variants					
				IN-1K	IN-Real	IN-V2	IN-A	IN-R	IN-S
CLIP-L [11]	Linear	336	-	85.3	88.8	75.8	-	-	-
SigLIP-L [6]	Attentive	224	-	86.5	-	-	-	-	-
AIMv2-L [61]	Attentive	224	-	86.6	-	-	-	-	-
UniViTAR-0.3B	Linear	Native	83.0	87.6	90.3	79.5	**84.1**	90.6	66.0
UniViTAR-0.3B	Attentive	Native	**83.3**	**87.7**	**90.5**	**79.8**	83.8	**91.1**	**66.8**
CLIP-H [11]	Linear	224	-	84.4	88.4	75.5	-	-	-
†DFN-H [47]	Linear	378	81.6	87.3	90.4	78.8	74.8	90.3	68.3
SigLIP-SO [6]	Attentive	384	-	87.3	-	-	-	-	-
AIMv2-H [61]	Attentive	224	-	87.5	-	-	-	-	-
UniViTAR-0.6B	Linear	Native	84.4	88.2	90.6	80.6	87.1	92.0	68.0
UniViTAR-0.6B	Attentive	Native	**84.8**	**88.3**	**90.7**	**81.0**	**87.3**	**92.5**	**68.8**
OpenCLIP-G [45]	Linear	224	78.5	86.2	89.4	77.2	63.8	87.8	66.4
DINOv2-g [13]	Linear	224	78.6	86.5	89.6	78.4	75.9	78.8	62.5
EVA01-g [49]	Linear	224	79.1	86.5	89.3	77.4	70.5	87.7	63.1
AIMv2-1B [61]	Attentive	224	-	88.1	-	-	-	-	-
InternViT-6B [12]	Linear	224	82.5	88.2	90.4	79.9	77.5	89.8	69.1
EVA-8B [49]	Linear	224	-	88.5	-	-	-	-	-
UniViTAR-1B	Linear	Native	**86.0**	88.9	90.8	81.5	90.1	**94.0**	**70.7**
UniViTAR-1B	Attentive	Native	**86.0**	**89.2**	**91.0**	**81.7**	**90.1**	93.6	70.6

3.5 Results on Dense prediction.

In this section, we evaluate the dense prediction performance of our UniViTAR family by transferring to semantic segmentation. Following [12, 63], we fine-tune a decoder with freezing backbones under two different structures, *i.e.*, Linear and UperNet. Linear decoder transforms the dimension of one single layer visual feature to number of semantic classes, while the UperNet decoder employs PPM and FPN to integrates multi-scale features. Experiments are conducted on the ADE20K [64] dataset. In terms of data preprocessing, we employed the same fixed-resolution input and data augmentation strategies as those used in InternViT [12]. Corresponding results are shown in Table 6. We can observe a performance gap between these two types of decoder, this can be understand that UperNet has significantly more trainable parameters than Linear decoder. Taking UniViTAR-0.6B as an example, Linear decoder has a parameter count of 0.2M, whereas UperNet contains approximately 200M parameters. Notably, our UniViTAR Family demonstrates an obvious performance advantage compared with existing state-of-the-art vision encoders. Under the setting of Linear decoder, our UniViTAR-1B achieves a performance of 45.4 mIoU, which is $+6.1$ points over OpenCLIP-G [45] and $+10.8$ points over ViT-22B [10]. In the case of UperNet decoder, our UniViTAR-1B reaches 56.2 mIoU, also surpassing larger parameter-scale model like InternViT-6B [12].

Table 6: **Evaluation of semantic segmentation on ADE20k dataset with frozen backbones.**

Method	CropSize	mIoU^{Linear}	mIoU^{UperNet}
CLIP-L [11]	-	39.0	-
SigLIP-SO [6]	-	40.8	-
†DFN-H [47]	-	41.3	-
OpenCLIP-G [45]	512^2	39.3	
InternViT-6B [12]	504^2	**47.2**	54.9
ViT-22B [10]	504^2	34.6	52.7
UniViTAR-0.3B	504^2	40.7	54.6
UniViTAR-0.6B	504^2	42.9	55.1
UniViTAR-1B	504^2	45.4	**56.2**

3.6 Results on Multimodal Understarding

Evaluation Setup. To assess the potential of multimodal understanding, we employ a dual-stage training paradigm, similar to common practices [19, 20]. In the pretraining stage, we train the projector with a learning rate of $1e^{-3}$ using a merged 2.5M dataset comprised of LLaVA-CC3M-Pretrain [17], ALLaVA-Caption [19], ShareGPT4V-PT [20]. In the fine-tuning stage, we unfreeze the whole model, and train it with a learning rate of $1e^{-5}$, using the high-quality instrution-tuning

dataset LLaVA1.5-Finetune [65]. Note that the native-resolution strategy of *Boundary Markers* and *Line Anchors* are only applied in the fine-tuning stage. All evaluations are conducted using VLMEvalKit [66], assessing performance across 16 popular benchmarks, including GQA [67], DocVQA [68], InfoVQA [69], ScienceQA [70], TextVQA [71], VizWiz [72], OCRVQA [73], OCR-Bench [74], MME [75], MMMU [76], SEEDBench_IMG [77], MathVista_MINI [78], AI2D [79], HallusionBench [80], POPE [81], HRBench4K [82].

Results Comparison and Analysis. As illustrated in Table 7, under exactly the same training data and training strategy, the proposed UniViTAR surpasses various state-of-the-art vision encoders [47, 61, 6] on numerous multimodal understanding benchmarks. Notably, UniViTAR demonstrates exceptional capabilities in scenarios involving dense information, such as document parsing [68], graphic parsing [69], and high-resolution tasks [82]. We argue that the native resolution plays a crucial role in achieving outstanding performance in these areas, ensuring minimal loss of image information. We also assess the effectiveness of the proposed strategy of *Boundary Markers* and *Line Anchors* on 0.6B model size, as demonstrated in Table 7, which highlights their impact.

Table 7: **Evaluation of multimodal understanding on various vision-language benchmarks**. Note the superscript $\triangle$ represents model with *Boundary Markers* and *Line Anchors*.

Benchmarks	**SigLIP-L** [6]	**DFN-H** [47]	**AIMv2-H** [61]	**SigLIP-SO** [6]	**UniViTAR**			
					0.3B$^{\triangle}$	**0.6B**	**0.6B**$^{\triangle}$	**1B**$^{\triangle}$
Resolution	378	378	448	384	**Native**			
$\text{GQA}_{\text{TestDev}}$	61.5	60.6	61.5	61.0	60.8	58.2	60.3	61.2
$\text{DocVQA}_{\text{VAL}}$	30.8	25.9	36.2	32.0	47.7	46.3	48.2	47.0
$\text{InfoVQA}_{\text{VAL}}$	22.7	22.1	25.8	23.2	27.8	28.0	28.5	27.5
$\text{ScienceQA}_{\text{VAL}}$	63.6	62.7	64.8	66.4	64.5	65.0	63.6	65.3
$\text{TextVQA}_{\text{VAL}}$	48.0	41.7	53.2	50.9	50.8	52.0	50.7	52.0
VizWiz	30.5	28.5	30.3	30.8	29.1	29.8	29.4	29.3
OCRVQA	31.2	32.0	31.0	30.9	32.2	31.6	32.2	32.1
OCRBench	35.2	30.6	22.4	36.0	33.6	37.0	36.9	36.4
MME	59.3	62.6	59.8	60.0	57.9	58.6	59.0	60.7
MMMU_{VAL}	35.9	34.2	37.1	35.4	36.1	38.7	36.6	37.0
SEEDBench_IMG	70.0	70.3	70.9	71.2	68.2	67.8	68.0	69.3
MathVista_MINI	28.6	29.9	30.1	29.6	27.5	27.9	28.5	28.7
$\text{AI2D}_{\text{Test}}$	60.5	58.3	60.4	60.6	58.0	57.7	58.7	59.3
HallusionBench	56.8	56.6	53.9	54.8	57.0	55.2	57.8	54.1
POPE	87.2	88.0	85.4	87.7	87.1	88.1	87.9	88.1
HRBench4K	39.9	39.5	44.5	45.0	44.1	43.6	46.1	46.4
Average	47.6	46.5	48.0	48.5	48.9	49.1	49.5	49.6

3.7 Ablation Study

3.7.1 Robustness verification of resolution mode.

This section analyzes the influence of three resolution modes: fixed resolution, native aspect ratio, and native resolution in Figure 3. For fixed resolution mode, we resize the shorter edge of each image to a predefined size S and apply *CenterCrop* to ensure the sequence length strictly equals $(S/14)^2$, where 14 represents the patch size. Increasing S proportionally extends the sequence length. In native aspect ratio mode, we scale images while preserving their original aspect-ratio, ensuring that $wh/14^2$ approximates the target sequence length. We evaluate 12 sequence lengths ranging from 256 to 16K tokens with our 0.6B model. The experimental results reveal three key findings: *1)* performance initially improves then declines with increasing sequence length under both fixed and native aspect ratio modes, peaking at 1024$\sim$4096 tokens. *2)* native aspect ratio consistently outperforms fixed resolution, indicating that preserving original aspect ratios retains better image information. *3)* native aspect ratio occasionally surpasses native resolution performance at certain lengths.

3.7.2 Verification of the effectiveness of training strategies.

As introduced in the method, we categorize the training strategy into four distinct stages, based on resolution modes (fixed or native), visual data modalities (image or video), and model training parameters (trainable or freeze). As shown in Figure 4, for *zero-shot image classification* (left), we show that S1, S2, and S3 exhibit progressive performance improvements, while S4 maintains comparable accuracy despite incorporating alternating training. In contrast, for *zero-shot video classification* (right), S1 and S2 show minimal performance variation, with dynamic-resolution

Figure 3: **Performance comparison of different resolution modes as the length of the vision sequence increases**. The black dashed line shows the performance when using native resolution.

training in S3 significantly boosting video capabilities, followed by further enhancements in S4 through image-video training. This demonstrates that dynamic-resolution training enables model to process more native visual sequences, while the final unified training stage equips the model with generalized capabilities for handling diverse visual modalities. Notably, these findings remain highly consistent across the UniViTAR-0.3B/0.6B/1B model family.

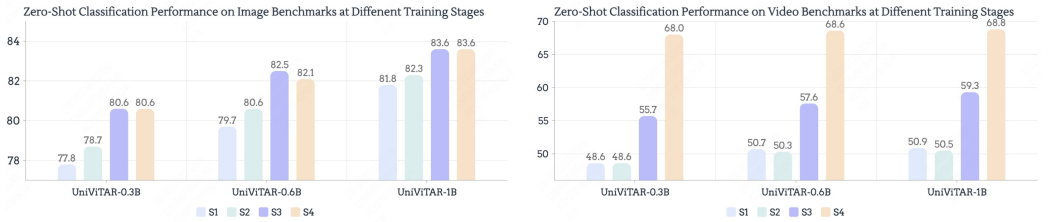


Figure 4: **Average performance improvement illustration across different training stages**.

3.8 Verification of the effectiveness of hybrid training with DINOv2.

To investigate the performance benefits of incorporating DINOv2 as a distillation branch in Stage-1 training, we performed a comprehensive empirical study using the UniViTAR-0.3B model trained from scratch on 3B image-text pairs. Checkpoints were evaluated at regular intervals throughout training in Table 8. Our experiments revealed that DINOv2 distillation significantly accelerates early-stage convergence, after only 0.1B samples, distillation improved zero-shot ImageNet-1K performance by 17.3 points. Although this gain gradually diminishes as training progresses, it remains observable at later stages. More importantly, the final model trained with DINOv2 distillation achieves an average improvement of 2.1 points across six zero-shot classification benchmarks compared to the baseline without distillation (Table 9). These results demonstrate that DINOv2 distillation not only speeds up early convergence but also enhances the final model performance.

Table 8: **The performance gains of hybrid training on ImageNet-1K as data increases**.

Model	DINOv2	0.1B	0.5B	1.0B	1.5B	2.0B	2.5B	3.0B
UniViTAR-0.3B	No	26.88	63.71	67.80	69.75	72.32	74.88	75.72
UniViTAR-0.3B	Yes	44.18	68.15	71.24	73.17	74.71	76.40	77.33
Δ	-	17.30	4.44	3.44	3.42	2.39	1.52	1.61

Table 9: **Zero-shot classification performance of hybrid training with DINOv2**.

Model	DINOv2	Avg.	IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net
UniViTAR-0.3B	No	70.73	75.72	58.76	87.98	68.22	62.95	70.78
UniViTAR-0.3B	Yes	72.84	77.33	63.55	89.81	70.40	65.65	70.31
Δ	-	2.11	1.61	4.79	1.83	2.18	2.70	-0.47

4 Conclusion

In this work, we introduce UniViTAR, a family of homogeneous vision foundation models tailored for unified visual modality and native-resolution scenarios in the era of multimodal. By integrating advanced architectural upgrades, resolution curriculum learning, visual feature distillation, and inter-batch modality adaptation, UniViTAR achieves significant improvements across diverse tasks, spanning image/video zero-shot classification/retrieval, dense prediction accuracy, and vision-language model transfer performance. Notably, all models are trained exclusively on public-accessible datasets, where we observe consistent performance gains with parameter scaling from 0.3B to 1.4B. We hope that our UniViTAR offers the community a versatile framework for advancing multimodal research.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023.

[3] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.

[4] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[6] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[7] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[9] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[10] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[14] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[15] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[16] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[18] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024.

[20] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[21] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

[22] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022.

[24] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[25] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021.

[26] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[27] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[28] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[29] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[30] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*, pages 3505–3506, 2020.

[31] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[32] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[33] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 35:16344–16359, 2022.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[35] 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*, pages 15262–15271, 2021.

[36] 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. *ICCV*, 2021.

[37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019.

[38] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

[39] 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. *NeurIPS*, 32, 2019.

[40] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022.

[41] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[42] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

[43] 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 *ECCV*, pages 740–755, 2014.

[44] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[45] 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. Zenodo. Version 0.1. `https://doi.org/10.5281/zenodo.5143773`, 2021. DOI: 10.5281/zenodo.5143773.

[46] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

[47] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[48] Xianhang Li, Zeyu Wang, and Cihang Xie. Clipa-v2: Scaling clip training with 81.1% zero-shot imagenet accuracy within $10000 budget; an extra $4000 unlocks 81.8% accuracy. *arXiv preprint arXiv:2306.15658*, 2023.

[49] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[50] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[51] 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.

[52] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[53] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[54] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[55] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[56] Mamshad Nayeem Rizve, Fan Fei, Jayakrishnan Unnikrishnan, Son Tran, Benjamin Z Yao, Belinda Zeng, Mubarak Shah, and Trishul Chilimbi. Vidla: Video-language alignment at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14043–14055, 2024.

[57] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022.

[58] Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024.

[59] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051, 2022.

[60] Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shu-Tao Xia, and Ying Shan. Tvtsv2: Learning out-of-the-box spatiotemporal visual representations at scale. *arXiv preprint arXiv:2305.14173*, 2023.

[61] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024.

[62] 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.

[63] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, pages 7480–7512, 2023.

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.

[65] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[66] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

[67] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[68] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[69] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022.

[70] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[71] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.

[72] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018.

[73] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[74] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[75] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[76] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[77] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[78] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[79] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[80] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[81] Kun Zhou Jinpeng Wang Wayne Xin Zhao Yifan Li, Yifan Du and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[82] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint*, 2024.

[83] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[84] David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. *arXiv preprint arXiv:2504.01017*, 2025.

[85] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[86] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *arXiv preprint arXiv:2205.12956*, 2022.

[87] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[88] Qihang Fan, Quanzeng You, Xiaotian Han, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Vitar: Vision transformer with any resolution. *arXiv preprint arXiv:2403.18361*, 2024.

[89] 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*, pages 770–778, 2016.

[90] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[91] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[92] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186. IEEE Computer Society, 2023.

[93] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[94] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[95] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[96] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[97] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[98] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1236–1248, 2024.

[99] openai. Language models are unsupervised multitask learners. `https://openai.com/index/hello-gpt-4o/`, 2024.

[100] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[101] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.

[102] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

[103] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024.

[104] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[105] Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, and Jie Zhou. Points1.5: Building a vision-language model towards real world applications, 2024.

[106] 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, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. Our main contributions are also detailed in Sec. 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We analyzed the performance gap with the pure video model in the experiment section. In addition, the reduction in inference efficiency caused by native resolution is also mentioned in the article.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Each experiment in the paper provides a detailed description of its setting, referring to Sec. 3 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly-accessable image-caption dataset. Once the blind review period is finished, we'll open-source code and model checkpoints.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The training hyperparameters are introduced in Table 11, and the test details are presented along with the experimental results in Sec. 3 and Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: This topic has not been reported with experimental statistical significance in other works.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the training resources in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the guidelines to ensure that our research strictly adheres to ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This research primarily focuses on fundamental network structures, with no potential social harm.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not foresee any high risk for misuse of work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have carefully reviewed the guidelines and ensured adheres to the standards.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing, human subjects, or other related risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing, human subjects, or other related risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We have carefully reviewed the guidelines and ensured adheres to the standards.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A The Overview of UniViTAR Pipeline

The proposed UniViTAR processes visual input at its native resolution, and also supports scaling the resolution down or up while maintaining the aspect ratio or resizing to certain square size to accommodate different application scenarios, such as higher computational efficiency or finer-grained visual details. By treating video inputs as temporally extended image sequences, the framework uniformly produces longer variable-length visual token sequences.

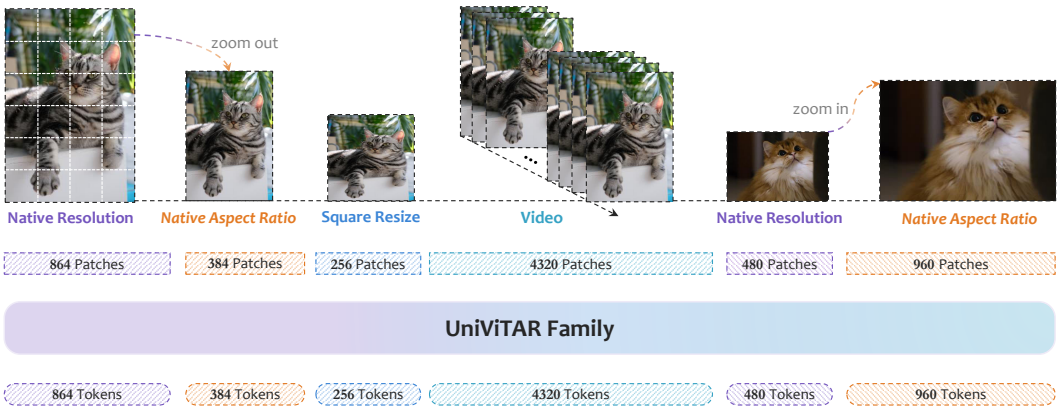


Figure 5: **The brief illustration of UniViTAR family pipeline**.

B More Details of Training Recipe

Data Details. We collect public accessible image-text pairs and build our Merged-1B dataset, which is composed of DataComp-1B [21], COYO [22], LAION-2B [23], LAION-400M [24], DFN-2B [22], CC12M [25] and CC3M [26]. Moreover, to further enhance the video feature extraction capabilities of UniViTAR, we meticulously constructed a dataset Merged-65M of roughly 65 million samples by randomly selecting video clips from three public accessible video datasets, *i.e.*, Panda-70M [27], WebVid-10M [28], and InternVid-10M-FLT [29]. We refer to the combined image and video data mentioned above as Merged-1.1B. The detailed data composition is summarized in the Table 10.

Table 10: **Details of the training data for UniViTAR**. Note that Merged-1B and Merged-65M correspond to image and video modality respectively.

Dataset	Source	Language	Samples	Total	Percentage	Used by
Merged-1B	DataComp-1B [21]	En	408M	1.08B	37.7%	Stage 1∼4
	COYO [22]	En	248M		22.9%	
	LAION-2B [23]	En	213M		19.7%	
	DFN-2B [47]	En	154M		14.3%	
	LAION-400M [24]	En	52.7M		4.9%	
	CC12M [25]	En	2.94M		0.3%	
	CC3M [26]	En	2.32M		0.2%	
Merged-65M	Panda-70M [27]	En	52.1M	65M	80.2%	Stage 4
	WebVid-10M [28]	En	6.53M		10.0%	
	InternVid-10M-FLT [29]	En	6.31M		9.8%	

Hyperparameter Details. The detailed hyperparameter configurations for each training stage are presented in the Table 11. As tabulated, we utilize a progressive reduction of the peak learning rate in correlation with increasing visual backbone scale to ensure optimal training stability. Notably, the learning rate of text branch in Stage 2 remains consistently one-tenth of the visual component throughout this phase. To enhance training efficiency, we integrated the DeepSpeed library [30] by employing ZeRO optimizer sharding [31], gradient checkpointing [32], and flash attention [33]. Note all experiments are conducted on H800 GPUs.

Table 11: **Detailed training hyperparameter of UniViTAR family**. Note that the symbol of $\rightarrow$ represents the peak learning rate and the minimum learning rate in the LR schedule.

	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**
Vision Encoder *Init.*	Xavier *init.* [83]	from Stage-1	from Stage-2	from Stage-3
Text Encoder *Init.*	LLama [12]	LLama [12]	from Stage-2	from Stage-3
Input Resolution	224×224	224×224	**Native**	**Native**
Token Range	256	256	$64 \sim 16$K	$64 \sim 16$K
Global Batch Size	32768	32768	32768	$\sim$26K(Image), $\sim$4K(Video)
Patch Dropout	0.5	0.0	0.5	0.5
Warmup Steps	2000	2000	2000	1000
Optimizer	AdamW	AdamW	AdamW	AdamW
LR Schedule	Cosine Decay	Cosine Decay	Cosine Decay	Cosine Decay
0.3B	$1e^{-3} \rightarrow 1e^{-6}$	$1e^{-5} \rightarrow 0$	$1e^{-5} \rightarrow 0$	$4e^{-6} \rightarrow 0$
0.6B	$1e^{-3} \rightarrow 1e^{-6}$	$1e^{-5} \rightarrow 0$	$1e^{-5} \rightarrow 0$	$4e^{-6} \rightarrow 0$
1B	$8e^{-4} \rightarrow 1e^{-7}$	$6e^{-6} \rightarrow 0$	$6e^{-6} \rightarrow 0$	$2e^{-6} \rightarrow 0$
Train Dataset	Merged-1B	Merged-1B	Merged-1B	Merged-1B, Merged-65M
Seen Samples	12B	1B	1B	0.6B

C More Experimental Results & Ablation Study

C.1 Verification of the effectiveness of image-video alternative strategy.

To validate the efficacy of the alternating image-video training strategy, we conducted initial experiments with 100M image-text pairs and 10M video-text pairs. Note that the image-to-video data ratio is approximately 10:1, consistent with the ratio used in stage 4 of the UniViTAR series. We trained a UniViTAR-0.3B model for 3 epochs, comparing mixed training and alternating training strategies. As shown in Table 12, the alternating training strategy outperforms the mixed strategy across key image and video benchmark metrics, demonstrating its effectiveness in enhancing visual representation learning. This performance gain can be attributed to the increased training difficulty arising from the unification of data modalities within each batch.

Table 12: **Zero-shot classification performance of image-video training strategy**.

Strategy	ImageNet-1K	ImageNet-A	K400	UCF101
Batch-Mixed	70.46	45.89	58.82	75.15
Batch-Alternative	71.25	48.60	61.01	77.66

C.2 Verification of the effectiveness of native resolution for video.

In this section, we conduct an ablation study to explore the role of native resolution in video data processing. We dynamically sample a maximum of 32 frames (denoted as F) for each video clip. For frames exceeding the sequence length limit, we resize them while preserving their native aspect ratio to a smaller resolution. We evaluate 15 maximum video sequence length, ranging from 1024 to 65,536, and test the zero-shot classification performance of UniViTAR-0.6B on the K400 dataset. Note that the minimum video sequence length is fixed to 576. As shown in Figure 6, the performance initially improves and then stabilizes as the sequence length limit increases, reaching a plateau at length 10,240. We attribute this to the fact that, with 32 sampled frames, a sequence length of 10,240 corresponds to a resolution of 490×256, enabling most frames in K400 to retain their native resolution during data processing. This finding underscores the importance of native resolution in enhancing video understanding capabilities.

C.3 Verification of the effectiveness of data scale.

From an intuitive perspective, data scale has a significant impact on the effectiveness of contrastive learning. In this section, we conduct cold-start experiments on UniViTAR-0.3B to confirm this view. For the experiment setup, seen samples is fixed at 1B. We respectively train the UniViTAR-0.3B for 1 epoch using Merged-1B and for 10 epochs using Merged-100M, which contains 100M image-text pairs that randomly sampled from Merged-1B. Result on zero-shot classification and retrieval is

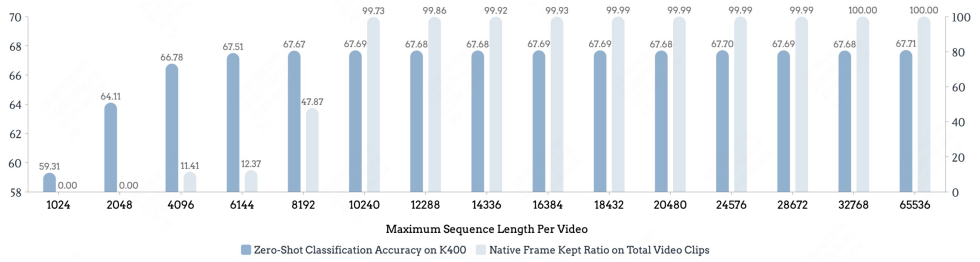


Figure 6: **Performance changes on the K400 dataset across varying sequence length limits.**

shown in Table 13. There is an observable trend where the performance improves as the dataset scale increases. With larger dataset scale, the model is exposed to a broader range of image-text pairs, facilitating a more comprehensive learning and understanding of the visual and linguistic space, thereby enhancing zero-shot performance.

Table 13: **Ablation results of UniViTAR-0.3B under varying data scale.**

Data	Seen Samples	Overall	ImageNet Variants						Overall	Flickr		COCO	
			IN-1K	IN-A	IN-R	IN-V2	IN-S	O-Net		T→I	I→T	T→I	I→T
Merged-100M	1B	60.8	69.7	39.6	79.0	61.6	54.8	60.2	67.0	73.4	88.3	44.2	62.0
Merged-1B	1B	64.2	71.7	45.7	82.3	64.3	57.3	63.9	68.9	74.9	90.7	46.0	63.8

C.4 Verification of the effectiveness of LLM scale.

To verify the model's effectiveness across language models of varying scales, we integrated UniViTAR-0.6B with progressively larger language backbones, specifically, Qwen2.5-1.5B, 3B, and 7B, within a multimodal large language model (MLLM) framework. As the model scale increases, our models achieve average scores of 48.7, 51.9, and 54.6, respectively, across 16 multimodal benchmarks. The experimental results demonstrate consistent scaling behavior, confirming UniViTAR's strong compatibility and performance potential when combined with larger language models. These findings support the conclusion that the UniViTAR architecture possesses promising scalability.

C.5 Comparative analyses of additional relevant visual foundation models.

To provide a comprehensive comparison with other prominent visual encoders, we conduct a systematic analysis of several relevant visual foundation models. (1) NaViT [3]: Quantitative evaluations in Table 14 on linear-probing classification tasks demonstrate that UniViTAR-0.3B exhibits clear advantages over NaViT. (2) FlexiViT [2]: Introduced in the main text, FlexiViT supports dynamic patch size to handle variable-resolution inputs. We include supplementary linear-probing results on ImageNet variants in Table 14 for direct comparison. (3) Web-DINO [84]: This self-supervised model shows that scaling data and parameters can approach CLIP-level performance; however, a noticeable gap remains relative to CLIP-based paradigms, as indicated by model size (7B vs. 0.3B) and benchmark performance. Preliminary comparisons are provided in the accompanying Table 14. (4) Cambrian-1 [85]: This vision-centric MLLM family uses a Mixture-of-Features (MoF) scheme over multiple visual encoders to reduce information loss. While effective, MoF introduces higher computational costs and integration complexity compared to unified models like UniViTAR.

Table 14: **Comparison of UniViTAR and other vision encoder on linear-probing classification**.

Model	Pretrain Data	Train Paradigm	IN-1K	IN-A	IN-Real	IN-V2	IN-S	IN-R
NaViT-L	JFT4B	Supervised Learning	76.0	65.5	-	-	-	-
FlexiViT-L	ImageNet-1K	Supervised Learning	86.1	34.1	90.0	76.7	-	41.2
Web-DINO-7B	MC-2B	Self-Supervised Learning	86.4	-	-	-	-	-
UniViTAR-0.3B	Merged-1B	Contrastive Learning	87.6	84.1	90.3	79.5	66.0	90.6

D Related Work

D.1 Flexible Vision Transformers

Vision Transformers have showcased impressive performance in numerous visual tasks, such as image classification [1], image language pre-training [11], etc. Those methods work only at a single, fixed resolution. Some works [86, 87] attempt to meet the need for fine-grained visual representation by adapting the model to a higher resolution during the fine-tuning stage. However, directly resizing the input to a fixed square resolution still limits their representation capacity in diverse visual scenarios. Recently, there are some works in vision transformers attempting to accommodate images with native resolutions with variable aspect ratios. ViTAR [88] proposes an adaptive token merger module to alleviate the constraints of fixed resolution and adapt to multi-resolution inputs. However, it still limited by a predefied number of tokens that the model ultimately aims to obtain. NaViT [3] introduces sample packing used in language modeling for handling variable sequence length of image patches. Meanwhile, it introduces a factorized positional embedding schema in vanilla ViT to support variable aspect ratios and extrapolate to unseen resolutions. Qwen2.5-VL [5] integrates an NaViT-like approach to support native input resolutions, and employs multiple training phases for adapting it to multimodal large languages models, including CLIP pre-training, vision-language alignment, etc.

D.2 Vision Foundation Models

The development of vision foundation models has progressed through distinct phases, beginning with supervised learning paradigm dominated by landmark architectures like ResNet [89] and ViT [1], which established performance benchmarks through reliance on labeled data. However, the field witnessed a paradigm shift with the rise of self-supervised learning, which circumvented annotation bottlenecks through three principal branches: contrastive learning frameworks like SimCLR [90] and MoCo [91], masked image modeling methods such as BEiT [92] and MAE [93], and self-distillation techniques including BYOL [94] and DINO [95]. Recently, language-supervised contrastive pre-training has emerged as a transformative paradigm, exemplified by CLIP [96], which aligns multi-modal embeddings through noise-robust contrastive objectives, enabling zero-shot task generalization. This approach has been further refined in works like SigLIP [6], which employs a more efficient sigmoid-base loss function while preserving cross-modal transfer capabilities. Besides images, a robust visual foundation model with effective video alignment capabilities serves as another critical building block. The existing strategies for training such models can be classified into three main paradigms: training on video-only data [56, 58, 15, 97], utilizing multimodal data encompassing both video and image [15, 87, 4, 5], and incorporating multimodal data that integrate video, images, audio and other modalities [16, 98]. VideoPrism [58] employs a two-stage video-only pretraining strategy: contrastive learning followed by token distillation, yet lacks image understanding. VidLA [56] adapts CLIP [96] via spatio-temporal attention on video-text data. InternVideo [15] combines masked video modeling with alternating video/image-text pretraining, enhanced by cross-modal attention, while InternVideo2 [16] extends this framework with audio/speech modalities for multimodal alignment.

D.3 Multimodal Large Language Models

Recently, multimodal large language models (MLLMs) have witnessed significant advancements and rapid development [99, 17, 100, 87, 5, 4, 12, 101, 102]. As a critical modality in MLLMs, visual input encounters inherent limitations when relying on conventional ViT with fixed resolutions, which may induce shape distortions, content blurring, and suboptimal handling of images/videos with diverse aspect ratios, high resolutions, or dynamic frame rates. To mitigate these challenges, the field has converged on two principal technical directions: 1) The tiling-based paradigm, as adopted by models like [103, 100, 12, 104], decomposes ultra-high-resolution inputs into a varied number of fix-resolution tiles, and each tile is processed by a fixed-resolution vision encoder. As such, it enables MLLMs adaptivity to dynamic-resolution images without padding or shape-distorting resizing. However, the tile limits the model's ability to capture spatial information across different tiles and the primary subjects of the images are often fragmented, leading to the loss of spatial relationships and quantitative information. 2) native-resolution methodology, exemplified by models such as [4, 105, 106], attempts to circumvent the limitations of the tiling-based paradigm by using native resolution input. However, they typically employ a pretrained fixed-resolution vision transformer as vision encoder, which leads to additional costs associated with adapting the ViT's distribution.