PolyVision: Conditional Visual Scaling Via Dynamic Expert Routing For Vision-Centric MLLMs

Tiehan Fan **Chen Zhao** **Nikai Du** **Zili Yi** **Jian Yang** **Ying Tai**[†]
State Key Laboratory of Novel Software Technology, Nanjing University, China

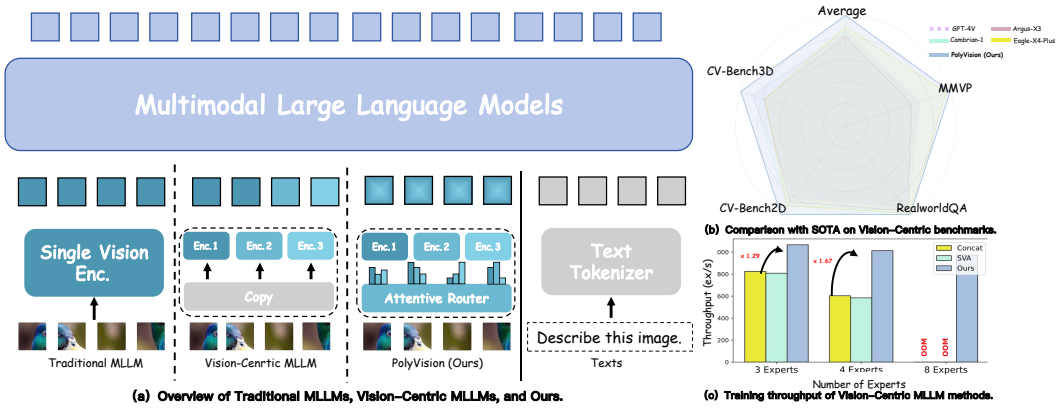


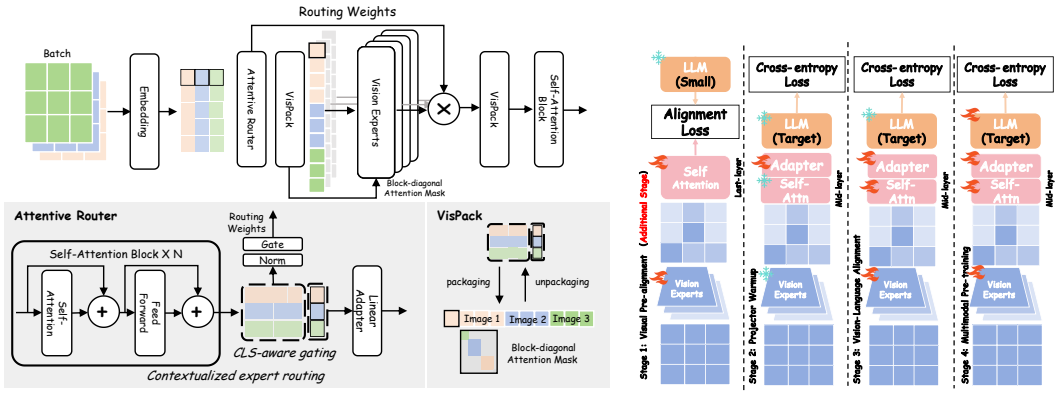
Figure 1: **Overview of the comparison between PolyVision and other paradigms.** **(a)** Prior VC-MLLMs scale visual capacity by adding experts, whereas PolyVision scales visual capacity by allocating experts through an attentive router. **(b)** Our method is highly competitive with SOTA. **(c)** Compared with other vision-centric paradigms, our method has significant advantages in memory and training speed. Concat represents the architecture used by Eagle and Argus, while SVA represents Cambrian-1. All methods use the same GPUs, effective batch size of 18K, and micro-batch size. For simplicity, we test with homogeneous DINOv2-Large experts.

Abstract

Multimodal large language models have achieved remarkable progress, yet their language-dominant designs constrain visual perception and reasoning. Vision-centric multimodal large language models aim to restore visual understanding as the structural foundation, but existing systems rely on static visual expert fusion, leading to redundant computation, weak adaptivity, and limited multimodal transferability. Prior VC-MLLMs scale visual capacity by **adding experts**, whereas `PolyVision` scales visual capacity by **allocating experts**. We introduce `PolyVision`, a vision-centric framework for conditional visual scaling in which dynamic visual expert routing assigns image patches to heterogeneous visual encoders. An **Attentive Router** selectively activates relevant experts for each patch, while a compact **VisPack** module executes routed representations efficiently and preserves example independence. Together with additional visual pre-alignment, this design turns heterogeneous visual scaling into a practical conditional-computation process. Built on the `Qwen3` backbone, `PolyVision` achieves superior efficiency and scalability on multimodal benchmarks. Experiments show that adaptive expert routing strengthens visual specialization, clarifies practical properties of visual scaling, and improves visual understanding at a favorable trade-off between efficiency and performance.

[†]Correspondence to: Ying Tai

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Overview of `PolyVision`. (b) PolyVision training pipeline.

Figure 2: **(a)** We realize conditional visual scaling through the Attentive Router and VisPack. Unlike existing vision-centric MLLMs, we dynamically allocate heterogeneous visual experts rather than uniformly activating static experts. **(b)** The framework is trained in four progressive stages, gradually activating more components until full joint optimization.

1 Introduction

Large language models (LLMs) have demonstrated remarkable generalization across a broad range of tasks [23, 37, 58, 15], prompting intensive interest in their extension to multimodal settings. This momentum has produced many multimodal large language models (MLLMs) that typically rely on well-trained visual encoders (either trained on image–text pairs [41, 53, 54] or via self-supervised visual pretraining [38, 34]). These visual encoders convert patches into visual tokens that are then projected into a language embedding space through a cross-modal connector. Yet multimodal learning remains a strict superset of language understanding, and current MLLMs still depend overwhelmingly on the linguistic tower while treating vision as a loosely coupled "plugin." This imbalance amplifies the limitations of visual encoders and constrains downstream multimodal performance [51, 52].

A promising alternative, vision-centric multimodal large language models (VC-MLLMs), re-center vision as the structural core of multimodal understanding, treating visual perception and representation as the primary drivers of model design and evaluation [51, 29, 24, 20, 16]. In this paradigm, a collection of visual encoders act as "visual experts," whose outputs the language model integrates, localizes, and reasons over. This design redefines multimodal integration as a process of visual specialization and alignment rather than mere feature fusion. However, current VC-MLLMs remain constrained by *static expert scaling*: the number and usage of visual experts are typically fixed, limiting both efficiency and adaptability.

Limitation of static expert inputs scaling. Existing VC-MLLMs often expand their visual capacity through static concatenation or uniform fusion of multiple visual encoders. Such strategies scale feature dimensionality or the number of visual tokens roughly linearly with the number of experts, increasing training and inference cost. More critically, static fusion forces every expert to process every patch, leading to redundant computation, representational overlap, and knowledge interference. This rigidity yields diminishing returns as expert count grows (Table 1), especially when self-supervised experts remain only weakly aligned with language. Together, these issues expose the limits of static expert scaling, **larger ensembles do not necessarily translate into more capable or efficient VC-MLLMs** [45] (Table 1).

From static fusion to conditional visual scaling. To address these limitations, we propose `PolyVision`, a vision-centric framework for *conditional visual scaling* in multimodal LLMs. Prior VC-MLLMs scale visual capacity by adding experts, whereas `PolyVision` scales visual capacity by allocating experts. Instead of statically concatenating experts or forcing every expert to process every patch, `PolyVision` introduces an *Attentive Router* that dynamically assigns image patches to heterogeneous visual experts according to their perceptual affinities. This adaptive routing mechanism selectively activates only relevant experts for each patch, maintaining computational efficiency while enhancing specialization. To prevent token explosion, we further develop a *VisPack* module that compactly packages each expert's latent-space inputs for parallelized processing, pre-

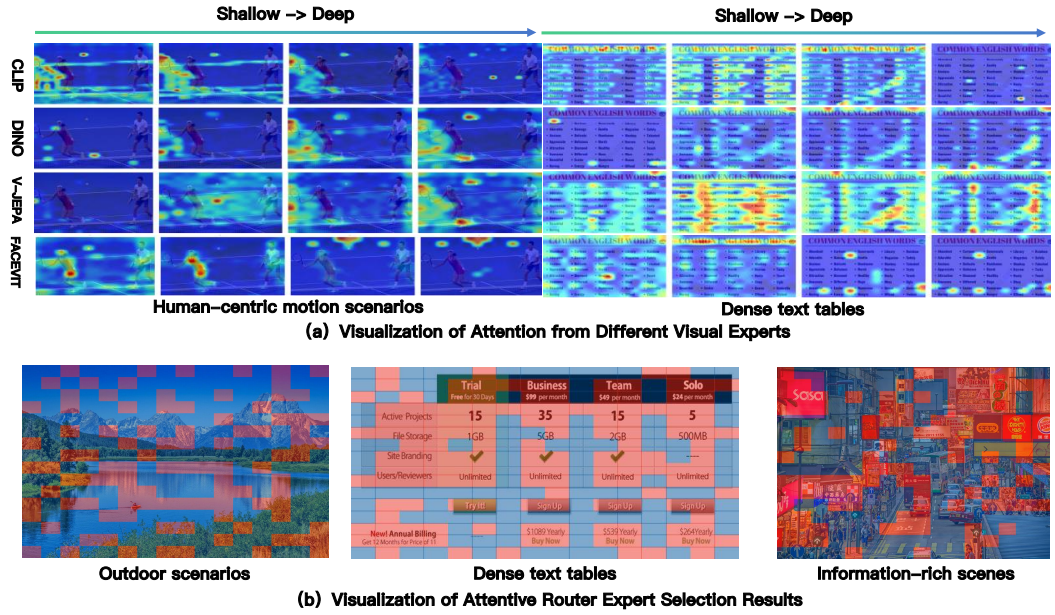


Figure 3: **Visualization of visual expert behaviors and routing preferences. (a)** Different visual experts exhibit distinct preferences: CLIP focuses on text and semantic regions, DINO and V-JEPA [2] favor holistic structure, and FaceViT [39] emphasizes human cues. **(b)** Attentive Router assignments after training (without the zero expert). Red denotes CLIP-dominant patches and blue denotes DINO-dominant ones. CLIP dominates semantically rich regions, while outdoor scenes show alternating CLIP/DINO dominance, indicating adaptive expert specialization.

serving expert independence without the high cost of naive feature concatenation. Rather than acting as isolated modules, these components jointly address a coupled challenge that existing systems do not fully resolve: expert allocation, efficient variable-length execution, and stable multimodal transfer across heterogeneous experts. We implement `PolyVision` on top of `Qwen3` [50], integrating diverse pretrained vision encoders via an additional visual pre-alignment and instruction-tuning pipeline that retains compatibility across expert types.

Towards better vision-centric multimodal LLMs. Extensive experiments further clarify several practical properties of conditional visual scaling, which we summarize here and analyze in detail in Supplementary Section S3.1. *(1) Diversity of experts matters.* Combining diverse visual experts with adaptive routing yields consistent improvements across multimodal benchmarks, while static or identical experts quickly saturate. *(2) Pre-alignment scaling benefits are bounded by model capacity.* Scaling visual pre-alignment data synergizes with adaptive routing, steadily improving representation capability. *(3) Vision scaling dominates.* Allocating the scaling budget to adaptive visual expansion produces significantly higher performance gains than increasing the language tower under the same compute envelope. These conclusions suggest that `PolyVision` offers a principled and efficient path toward expanding multimodal visual capacity. By transforming visual scaling into conditional expert allocation rather than a static design constraint, our framework enables the sustainable growth of vision-centric multimodal systems, **a faster and stronger vision-centric multimodal large language model** as shown in Figure 1. We provide an anonymized code package in the supplementary material for review and plan to release code and model weights with different expert combinations after the review process to facilitate future research.

2 Related Works

Multimodal large language models. Large Language Models (LLMs) have achieved strong generalization and generative performance across diverse tasks, motivating their extension to visual modalities and the emergence of Multimodal Large Language Models (MLLMs) [4, 12, 1, 6]. Early MLLMs predominantly adopt a single, language-supervised vision encoder (e.g., CLIP or SigLIP) that aligns visual features with text embeddings via image–text contrastive learning. While effective for global semantics, such supervision often overlooks fine-grained, pixel-level cues (e.g., precise

color [61, 62], spatial position [8, 9], and localized patterns [7, 10, 11]), constraining fine-grained perception and visual grounding in downstream tasks [20, 51, 52]. In contrast, purely self-supervised visual encoders may yield weaker instruction-following abilities [52]. These observations have catalyzed a vision-centric direction within MLLMs that broadens visual capacity by drawing on visual models pretrained under diverse objectives and datasets.

Vision-centric multimodal LLMs. A key insight in recent research is that the vision encoder forms a major bottleneck for MLLMs when trained only with global image–text supervision [20, 51, 16], leading to failures in fine-grained perception and spatial reasoning [52]. To address this, recent works integrate heterogeneous pretrained vision experts, language-supervised for global semantics and self-supervised (e.g., DINOv2) for local structure, sometimes extended with domain-specific experts such as depth, segmentation, and OCR [29, 32]. Their features are fused through concatenation or cross-attention-based resampling [24, 51], consistently improving performance on perception-heavy benchmarks [52]. Yet current vision-centric designs still face scaling challenges: concatenation-based fusion increases token and channel costs, connector-stage fusion tightly couples vision–language optimization without true pixel-level specialization, and many experts remain poorly language aligned. These issues call for more principled, scalable fusion and alignment strategies built from the patch level upward. We therefore present `PolyVision`, a vision-centric MLLM that realizes conditional visual scaling through adaptive patch-wise expert allocation and efficient expert execution (Figure 2a).

3 Method

3.1 Attentive Router

Motivation. Prior studies [19, 13, 40] show that visual experts trained on heterogeneous datasets or objectives develop complementary capabilities and attend to distinct spatial regions or semantic cues (Figure 3 (a)). The motivation is to realize conditional visual scaling while preserving each expert's domain-specific knowledge and activating it only where it is most useful, rather than enforcing homogeneous processing across all inputs. In this setting, conditional visual scaling means that visual capacity is expanded through input-dependent expert allocation rather than uniformly increasing computation for every patch. Instead of treating additional experts as statically fused capacity, the model scales vision by selectively invoking complementary experts according to image content and context.

Mixture-of-Experts (MoE) architectures [43, 17] provide a natural sparse-routing template and let us leverage mechanisms such as zero-expert [21] and load balancing [64]. However, conventional MoE routers are typically simple MLPs that depend only on local embeddings and ignore global dependencies, which often yields unstable routing in vision tasks. `PolyVision` is therefore not a straightforward application of standard MoE: standard MoE assumes homogeneous experts in a shared hidden space, whereas our setting must allocate image patches across heterogeneous pretrained visual encoders with different perceptual biases, language alignment, and execution behavior.

Contextualized expert routing. We enhance routing reliability using a lightweight residual self-attention module that contextualizes token representations before expert assignment. For each example b, let $\hat{\mathbf{X}}_b = [\, c_b;\, x_{b,1}; \ldots; x_{b,n_b}\,] \in \mathbb{R}^{(n_b+1)\times d}$ be the concatenation of the CLS and patch tokens. The Attentive Router first computes enriched representations:

$$\mathbf{H}_b = \mathrm{SelfAttn}\big(\,\mathrm{LN}(\hat{\mathbf{X}}_b)\big) + \hat{\mathbf{X}}_b, \tag{1}$$

where $\mathrm{SelfAttn}(\cdot)$ is a multi-head self-attention layer with residual connection. The expert assignment probabilities are then computed as

$$p(e_i \mid x_{b,j}) = \mathrm{Softmax}\big(W_r \mathbf{H}_{b,j}\big)_i, \tag{2}$$

where $W_r \in \mathbb{R}^{E\times d}$ projects the contextualized embeddings to expert logits. Each token $x_{b,j}$ is routed to the top-k experts indexed by $\mathcal{K}_{b,j}$ according to these probabilities. Their outputs are fused as

$$y_{b,j} = \sum_{i\in\mathcal{K}_{b,j}} w_{b,j,i}\, f_{e_i}(x_{b,j}), \tag{3}$$

where $w_{b,j,i} = p(e_i \mid x_{b,j})$ is the original softmax probability, without renormalization over the selected experts, and f_{e_i} is the transformation function of expert e_i. The inclusion of contextualized states allows routing to consider both local features and cross-region semantics, yielding more stable and semantically coherent expert activations.

CLS-aware gating. To better align global semantics with the dominant local context, the CLS token is assigned to the expert most frequently activated across tokens in the current sample. Specifically, we first compute the activation counts

$$s_{b,i} = \sum_{j=1}^{n_b} \mathbf{1}[i \in \mathcal{K}_{b,j}], \quad i \in \{1, \ldots, E\}, \tag{4}$$

where $\mathcal{K}_{b,j}$ is the top-k expert index set for patch j of sample b. Let $\mathcal{T}$ be the set of trunk experts that accept CLS tokens. The dominant expert is $g_b = \arg\max_{i \in \mathcal{T}} s_{b,i}$, and the CLS representation h_b^{cls} is routed to this expert: $y_b^{\text{cls}} = f_{e_{g_b}}(c_b)$. This dominant-expert routing enables the CLS token to follow the primary semantic flow of the image, reinforcing consistency between global aggregation and the prevailing expert specialization pattern.

Time complexity analysis. We analyze the asymptotic cost of expert execution before VisPack, keeping only the dominant quadratic self-attention term and omitting lower-order operations such as layer normalization, linear projections, and routing logits. We assume comparable transformer experts, approximate load balance, no zero expert, and standard batch padding for activated experts.

Let B be the batch size, n the original visual sequence length, E the total number of experts, and $A \leq E$ the number of activated experts. Without routing, every expert processes the full sequence, so $C_{\text{full}} = \mathcal{O}(BEn^2)$.

Under routing, let $n_i \leq n$ be the padded sequence length processed by activated expert $i \in \mathcal{A}$. Then $C_{\text{routed}} = \sum_{i \in \mathcal{A}} \mathcal{O}\big(Bn_i^2\big)$. Under ideal load balancing, if $n_{\max} = \max_{i \in \mathcal{A}} n_i$, then $C_{\text{routed}} = \mathcal{O}\big(BAn_{\max}^2\big)$ with $n_{\max} \leq n$, and $C_{\text{full}}/C_{\text{routed}} = En^2/(An_{\max}^2)$.

Before VisPack, naive padding often keeps $n_{\max}$ on the same order as n. In that regime, $C_{\text{routed}} = \mathcal{O}(BAn^2)$ and the theoretical speedup is approximately E/A. Thus routing reduces the dominant cost by reducing the number of executed experts, while VisPack further removes padding waste from variable routed sequence lengths.

Toward execution efficiency. Although sparse routing improves specialization, it also produces variable token counts across experts. Standard batch padding to a uniform length $n_{\max}$ therefore wastes computation, since attention costs grow as $\mathcal{O}(BAn_{\max}^2)$. Inspired by NLP [25] and variable-resolution vision [14], we introduce **VisPack**, a variable-length token packing strategy that groups routed tokens from multiple samples into compact contiguous blocks. VisPack preserves per-sample isolation while absorbing variable sequence lengths into the batch dimension and eliminating redundant zero-padding.

3.2 VisPack: Variable-length Expert Execution

Given the routing mask $\mathcal{M} \in \{0,1\}^{E \times B \times n_b}$ and CLS routes $\{g_b\}$ from the Attentive Router, we group all tokens assigned to the same expert:

$$\tilde{\mathbf{X}}_{e_i} = \Big[\{c_b \mid g_b = i\}_{b=1}^{B};\ \{x_{b,j} \mid \mathcal{M}[i,b,j] = 1\}_{\substack{b=1,\ldots,B \\ j=1,\ldots,n_b}} \Big] \tag{5}$$

To ensure inter-sample independence, we apply a block-diagonal attention mask $\mathbf{M}_{e_i} \in \{0,1\}^{N_{e_i} \times N_{e_i}}$:

$$\mathbf{M}_{e_i}[u,v] = \begin{cases} 1, & \text{if tokens } u, v \text{ belong to the same sample} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Trunk experts ($i \in \mathcal{T}$) process both CLS and patch tokens, while non-trunk experts handle local patches exclusively. This produces compact expert-specific sequences that minimize redundancy while preserving semantic structure. Pseudo-code Algorithm S1 please see the appendix.

Table 1: **Performance of various vision-language models on different visual patterns in the MMVP-VLM benchmark [52].** We use symbols for visual patterns due to space limit: 🧭: Orientation and Direction, 🔍: Presence of Specific Features, 🔄: State and Condition, 🔢: Quantity and Count, 📍: Positional and Relational Context, 🎨: Color and Appearance, ⚙️: Structural and Physical Characteristics, 🅰️: Texts, 📷: Viewpoint and Perspective. The best results are in **bold**, and the second-best are underlined.

	Params (M)	Dataset size	🧭	🔍	🔄	🔢	📍	🎨	⚙️	🅰️	📷	MMVP Average
CLIP ViT-L/14 (OpenAI) [41]	427.6	400M	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
MetaCLIP ViT-L-14 [57]	427.6	1B	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [57]	986.1	1B	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [47]	1136.4	400M	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [47]	5044.9	2B	13.3	20.0	**66.7**	26.7	26.7	**66.7**	26.7	20.0	**33.3**	**33.3**
2 vision experts												
Concat	584.5	5M	13.3	20.0	6.7	20.0	6.7	26.7	6.7	13.3	6.7	13.3
SVA	622.8	5M	6.7	33.3	6.7	20.0	6.7	26.7	20.0	6.7	20.0	16.3
`PolyVision` (Ours)	613.7	5M	13.3	33.3	20.0	26.7	13.3	26.7	20.0	33.3	20.0	23.0
4 vision experts												
Concat	1464.1	5M	13.3	6.7	0	6.7	6.7	13.3	13.3	6.7	20.0	9.6
SVA	1214.0	5M	6.7	33.3	13.3	6.7	13.3	13.3	13.3	0.0	13.3	12.6
`PolyVision` (Ours)	1167.1	5M	**13.3**	**40**	20.0	13.3	20.0	46.7	**33.3**	**40.0**	13.3	26.7

Table 2: **Comparison with SOTA on MLLM benchmarks.** "Models†" are vision-centric multimodal large language models. The best results are in **bold**, and the second-best are underlined.

Model	**General**		**OCR & Chart**		**Vision-Centric**				
Method	MME^P	$SEED^I$	ChartQA	TextVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$	**Average**
GPT-4V	1409.4	69.1	78.5	78.0	50.0	61.4	64.3	73.8	62.4
Gemini-1.0 Pro	1496.6	70.7	-	-	-	-	-	-	-
Gemini-2.5 Pro	-	-	**81.3**	73.5	-	67.5	-	-	-
Grok-1.5	-	-	76.1	78.1	-	**68.7**	-	-	-
GPT-5.1-2025-11-13	-	-	-	-	**72.0**	65.8	73.3	87.4	74.6
Qwen3-VL [3]	**1725.2**	**78.8**	75.4	80.4	24.7	67.8	79.8	**93.0**	66.3
LLaVA-NeXT-8B [30]	1603.7	72.7	69.5	64.6	38.7	60.1	62.2	65.3	56.6
Mini-Gemini-HD-8B† [28]	1606.0	73.2	59.1	70.2	18.7	62.1	62.2	63.0	51.5
Eagle-X4-8B-Plus † [44]	1559.0	76.3	80.1	77.1	69.3	66.5	72.1	63.0	67.7
Argus-X3-8B† [33]	-	75.8	74.8	73.6	45.5	64.6	68.5	64.2	60.7
BRAVE† [24]	-	-	-	-	42.0	-	-	-	-
Cambrian† [51]									
+LLaMA3	1547.1	74.7	73.3	71.7	51.3	64.2	72.3	72.0	65.0
+Qwen3	1339.4	53.8	72.0	72.6	54.7	62.0	76.0	76.1	67.2
`PolyVision`†									
+LLaMA3	1638.7	74.8	74.6	73.1	68.0	62.5	78.1	75.3	71.0
+Qwen3	1657.3	**78.8**	76.5	**80.6**	**72.0**	67.3	**80.0**	81.9	**75.3**

3.3 Visual Pre-Alignment

Additional Visual Pre-Alignment has proved highly effective in VC-MLLMs [44]. Self-supervised visual encoders that are not pre-aligned with linguistic priors often fail to establish stable cross-modal correspondence during multimodal training and suffer degraded grounding performance [44, 51]. Building on recent advances [55, 59], we study two paradigms: the conventional CLIP text tower and the emerging LLM2VEC framework [60]. The former provides frozen linguistic anchors, while the latter uses an LLM to generate context-sensitive semantic embeddings (Figure 2b).

Through extensive experiments, we find that CLIP-style contrastive loss is the most stable and effective strategy. By contrast, adding patch-level alignment [5] often causes optimization instability, with the loss diverging to `NaN`, which we attribute to the granularity mismatch between patch-local visual embeddings and globally contextualized text features. Our final strategy therefore uses LLM2VEC with a residual MLP projection and contrastive learning between the fused visual [CLS] representation and the text embedding from the LLM's last token (see Section S3.2). Penultimate-layer patch hidden states are provided to the LLM as multimodal input [31], ensuring stable convergence and effective vision-language alignment (Figure 2b).

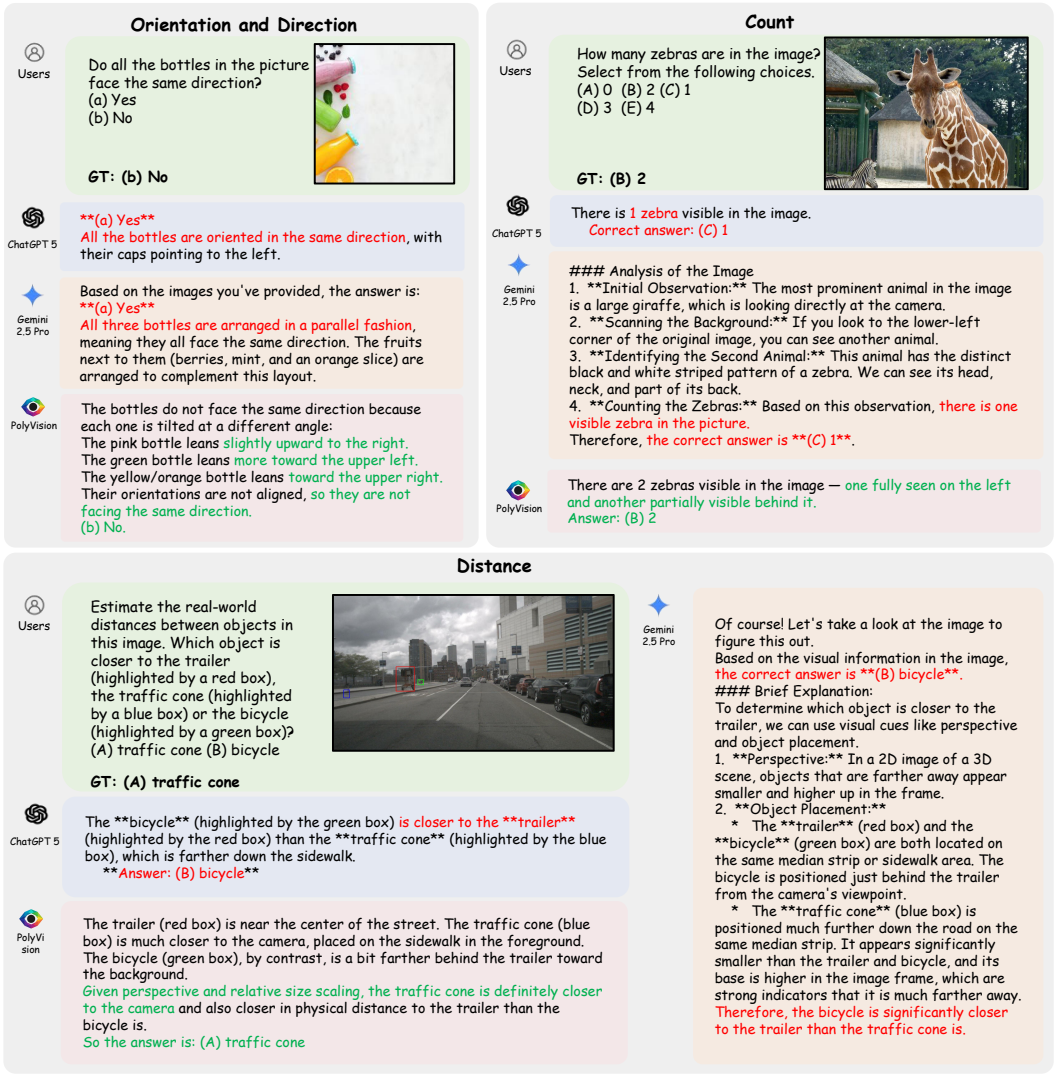


Figure 4: **Qualitative comparison with SOTA.** We compare ChatGPT-5 [36], Gemini 2.5 Pro [12], and `PolyVision` on three vision-centric tasks. `PolyVision` shows competitive fine-grained perception and captures task-relevant visual details. Examples come from MMVP (Orientation and Direction), CV-Bench-2D (Count), and CV-Bench-3D (Distance).

4 Experiments

4.1 Experimental Setup

We evaluate our model from two perspectives. (1) After visual pre-alignment, we compare the VLM with other representation models and vision-centric implementations using the same text backbone on MMVP-VLM [52], a manually annotated benchmark of visually challenging image pairs. (2) After full training, we compare `PolyVision` with VC-MLLMs and open-/closed-source MLLMs on the General [18, 26], OCR & Chart [35, 46], and Vision-Centric [52, 51, 56] benchmarks.

We use fixed expert combinations of {CLIP, DINOv2} and {CLIP, SigLIP, DINOv2, TrOCR}. Optimizer settings, expert selection, and learning-rate schedules are provided in the supplementary material and Table S1.

4.2 Comparison with SOTA

Introducing new experts does not affect efficiency. Figure 1 (c) compares throughput as expert count increases. By default, we activate two experts unless noted otherwise. As expert count grows,

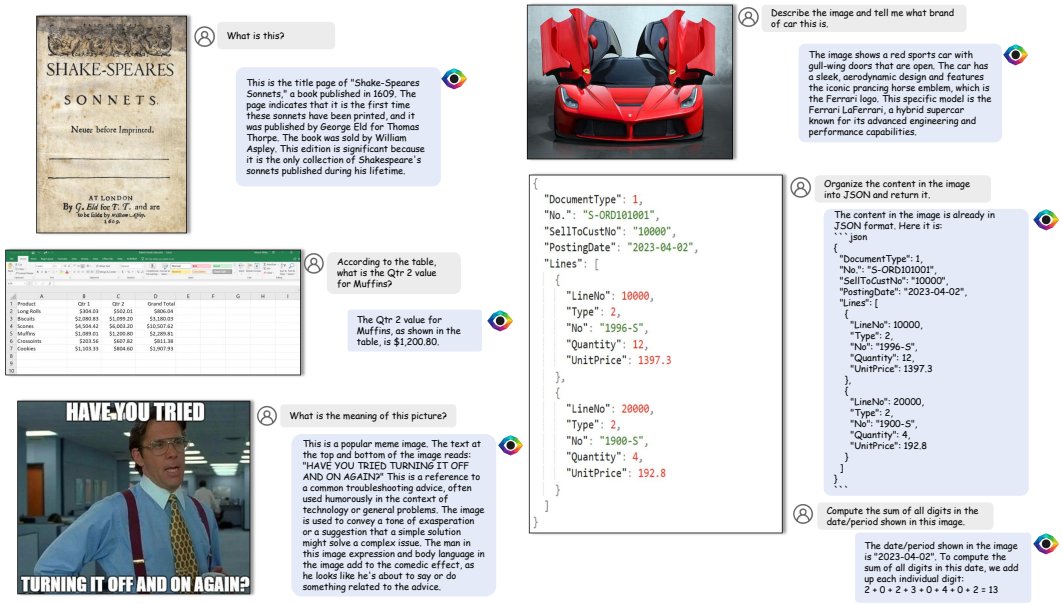


Figure 5: **More dialogue examples.** Our model supports text parsing, image understanding/annotation, multi-turn dialogue, and JSON-formatted output.

traditional methods run out of memory at the same micro batch size because VRAM usage scales sharply. In contrast, our method remains nearly unaffected because effective activation stays close to two experts.

Dynamic visual expert routing brings stable improvements. Table 1 compares representative VLMs. Under the same pre-alignment data scale and comparable parameter budgets, `PolyVision` achieves the highest MMVP Average, outperforming both SVA and naive Concat. We attribute this to spatial priors from heterogeneous experts, which fuse features more effectively than direct concatenation. After visual pre-alignment, `PolyVision` remains competitive with larger single-tower models pretrained on substantially larger image–text datasets.

These results also clarify a broader property of conditional visual scaling: greater heterogeneity sharply increases representation dimensionality, making image–text alignment harder under static fusion. This supports our view that *static expert scaling does not necessarily bring sustained gains.* We defer a detailed discussion of expert diversity, alignment data scale, and text-model scaling to Supplementary Section S3.1; the associated trends are visualized in Figure 6 (b).

Faster and stronger vision-centric multimodal large language models. Table 2 reports instruction-following results. Our model substantially outperforms vision-centric SOTA and remains competitive with traditional MLLMs. To disentangle architecture from backbone effects, we train Cambrian with Qwen3 and `PolyVision` with LLaMA-3. Even in this controlled comparison, `PolyVision` surpasses Cambrian, showing that the gains come from architecture rather than LLM strength. It also outperforms open-source Qwen3-VL on vision-centric benchmarks and exceeds commercial models such as GPT-5.1 on several tasks, confirming the benefit of dynamic expert routing. Performance gains also grow with data scale and backbone strength, while the dedicated OCR expert further improves general and OCR-related benchmarks.

Qualitative results. Qualitatively, our model outperforms strong commercial models on vision-centric tasks (Figure 4). The router weights also align with human preferences (Figure 3 (b)), and Figure 5 shows strong conversational ability beyond purely vision-centric tasks.

4.3 Ablation Study

Ablation on the attentive router. We replace the attentive router with a residual MLP block that performs only channel-wise transformations without inter-patch interaction. Although this variant adds parameters, it lacks spatial dependency modeling and cannot integrate global context across

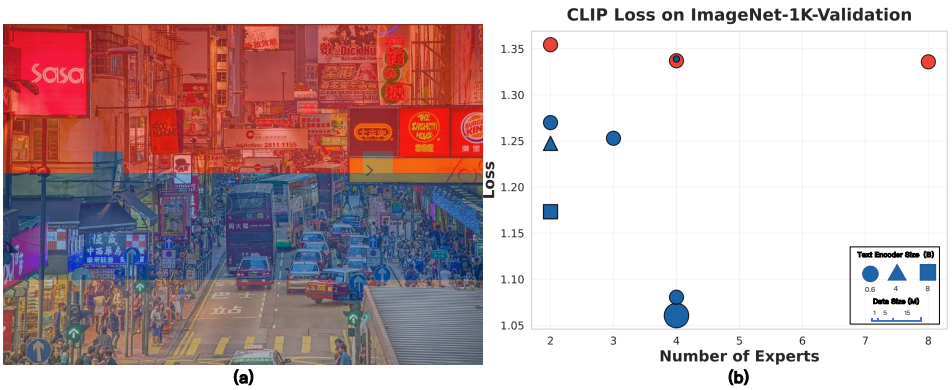


Figure 6: **(a) Gate weights of the model without the Attentive Router.** Red, Blue, and Gray denote the experts with the highest gating weights corresponding to CLIP, DINO, and the zero expert. **(b) Scaling behavior in PolyVision.** Where Red represents homogeneous experts and Blue represents heterogeneous experts.

Table 3: **Ablation on the attentive router.** Removing the attentive router causes a severe performance drop.

	Params (M)										MMVP Average
CLIP ViT-L/14 (OpenAI)	427.6	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
`PolyVision` (Ours)	613.7	13.3	33.3	20.0	26.7	13.3	26.7	20.0	33.3	20.0	23.0
w/o Attentive Router	639.9↑	6.7↓	6.7↓	6.7↓	6.7↓	13.3↓	13.3↓	13.3↓	20.0↓	6.7↓	10.4↓

Table 4: **Ablation on Visual Pre-Alignment.** Removing this stage degrades instruction-following performance.

Model	**General**		**OCR & Chart**		**Vision-Centric**				
`PolyVision-1B`	1143.7	52.3	77.9	77.4	34.0	52.7	48.3	51.1	46.5
w/o Visual Pre-Alignment	965.70↓	39.0↓	68.8↓	54.8↓	13.3↓	31.1↓	35.4↓	30.6↓	27.6↓

Table 5: **Ablation on VisPack, Top-K selection, and expert extension. Left:** Training throughput under different expert counts. **Right-Top:** MMVP under different top-K values with 5 experts. **Right-Bottom:** Adding a remote-sensing (R-S) expert trained on DescribeEarth-Dataset.

Experts	Method	Throughput (ex/s)	Speedup vs Concat
3	**Ours**	**1066 ± 10**	**1.29**
	w/o VisPack	1019 ± 5	1.23
4	**Ours**	**1013 ± 7**	**1.67**
	w/o VisPack	833 ± 2	1.38

Top-K	**MMVP Avg.**	**Std**	**Min/Max**
2	26.7	.054	.14/.28
3	26.7	.036	.13/.22
4	**28.1**	.024	.17/.22

Method	**MMVP**	**Recall@5**
w/o R-S expert	**26.7**	73.1
Only R-S expert	11.9	64.9
MoExtend	25.2	**78.1**

patches. As Table 3 shows, performance drops markedly because semantically coherent evidence becomes harder to aggregate across regions. Strong positional cues in routing are further visualized in Figure 6 (a).

Ablation on VisPack, Top-K selection, and Expert Extension. Removing VisPack and reverting to standard padding clearly reduces efficiency [14]. In Table 5, we keep the zero-expert mechanism for fairness. The slight acceleration of padding-only execution mainly comes from efficient zero-valued padding and the zero expert's ability to discard redundant tokens.

For top-K routing, we ablate MMVP across different values (Table 5, Right-Top). With five experts, performance remains stable over a reasonable range. We choose $K = 2$ by default because it balances efficiency and representation capacity; $K = 1$ collapses because expert diversity becomes insufficient, while $K = 3, 4$ yield comparable results with slightly lower variance.

Our architecture also supports dynamic expert extension through a unified initialization interface. We validate this by adding a remote-sensing expert trained on DescribeEarth-Dataset [27] using MoExtend [63]. As Table 5 (Right-Bottom) shows, although the R-S expert alone performs poorly on general benchmarks, MoExtend preserves original capabilities while significantly improving remote-sensing recall.

Ablation on visual pre-alignment. We remove the visual pre-alignment stage and directly train the 1B model from random initialization. This setting performs poorly across all benchmarks (Table 4), showing that pre-alignment is essential for stable early optimization. Without it, the embedding layers and router remain poorly conditioned, leading to gradient explosion, ineffective expert routing, and weak inter-modal mappings.

Training from scratch is especially difficult in the high-capacity heterogeneous-expert setting, where the router needs meaningful initialization to learn non-trivial routing distributions. This weakness is most evident on tasks relying on the OCR expert and complex visual features, making this stage even more necessary than in previous methods [51, 44].

5 Conclusion

In this work, we presented `PolyVision`, a scalable vision-centric multimodal architecture for conditional visual scaling via dynamic visual expert routing. By allocating raw image patches to specialized visual experts with an attention-based router, together with VisPack and additional visual pre-alignment, our method avoids token explosion while preserving expert independence and multimodal compatibility. `PolyVision` achieves strong results across multiple benchmarks and provides a useful framework for exploring heterogeneous visual representations in VC-MLLMs. Current vision-centric methods, including ours, still lack RoPE for multi-expert integration, which limits scaling to dynamically varying resolutions; we leave this to future work.

6 Acknowledge

This work was supported by Natural Science Foundation of China: No. 62406135, Natural Science Foundation of Jiangsu Province: BK20241198, the Gusu Innovation and Entrepreneur Leading Talents: No. ZXL2024362, and CCF-Tencent Rhino-Bird Open Research Fund.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[2] Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Mojtaba, Komeili, Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F.R., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning (2025), `https://arxiv.org/abs/2506.09985`

[3] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

[4] Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-vl technical report (2025), `https://arxiv.org/abs/2502.13923`

[5] Bica, I., Ilić, A., Bauer, M., Erdogan, G., Bošnjak, M., Kaplanis, C., Gritsenko, A.A., Minderer, M., Blundell, C., Pașcanu, R., et al.: Improving fine-grained understanding in image-text pre-training. In: Proceedings of the 41st International Conference on Machine Learning. pp. 3974–3995 (2024)

[6] Chen, Z., Wang, W., Cao, Y., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling (2025), `https://arxiv.org/abs/2412.05271`

[7] Chen, Z., Gao, R., Xiang, T.Z., Lin, F.: Diffusion model for camouflaged object detection. In: ECAI 2023: 26th European Conference on Artificial Intelligence, September 30–October 4, 2023, Kraków, Poland–Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023). pp. 445–452. SAGE Publications, 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP (2023)

[8] Chen, Z., Li, Y., Wang, H., Chen, Z., Jiang, Z., Li, J., Wang, Q., Yang, J., Tai, Y.: Ragd: Regional-aware diffusion model for text-to-image generation. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19331–19341. IEEE (2025)

[9] Chen, Z., Shi, T., Xu, P., Nan, K., Wang, Q., Yi, Z., Yang, J., Tai, Y.: Synvar: Synergizing spatial and semantic alignment in visual autoregressive model. In: European Conference on Computer Vision. pp. 201–218. Springer (2026)

[10] Chen, Z., Zhu, J., Chen, X., Zhang, J., Chen, J., Zeng, Z., Zhang, W., Wang, C., Yang, J., Tai, Y.: L2p: Unlocking latent potential for pixel generation. arXiv preprint arXiv:2605.12013 (2026)

[11] Chen, Z., Zhu, J., Chen, X., Zhang, J., Hu, X., Zhao, H., Wang, C., Yang, J., Tai, Y.: Dip: Taming diffusion models in pixel space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 36136–36146 (2026)

[12] Comanici, G., Bieber, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261

[13] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers (2024), https://arxiv.org/abs/2309.16588

[14] Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., 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)

[15] Du, Y., Fan, T., Nan, K., Xie, R., Zhou, P., Li, X., Yang, J., Yang, Z., Tai, Y.: Motionsight: Boosting fine-grained motion understanding in multimodal llms (2025), https://arxiv.org/abs/2506.01674

[16] Fan, X., Ji, T., Jiang, C., Li, S., Jin, S., Song, S., Wang, J., Hong, B., Chen, L., Zheng, G., Zhang, M., Huangcaishuang, Zheng, R., Xi, Z., Zhou, Y., Dou, S., Ye, J., Yan, H., Gui, T., Zhang, Q., Qiu, X., Huang, X., Wu, Z., Jiang, Y.G.: Poly-visual-expert vision-language models. In: First Conference on Language Modeling (2024), https://openreview.net/forum?id=7QaEO9WYMa

[17] Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)

[18] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), https://arxiv.org/abs/2306.13394

[19] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International conference on learning representations (2018)

[20] Jiang, D., Liu, Y., Liu, S., Zhao, J., Zhang, H., Gao, Z., Zhang, X., Li, J., Xiong, H.: From clip to dino: Visual encoders shout in multi-modal large language models (2024), https://arxiv.org/abs/2310.08825

[21] Jin, P., Zhu, B., Yuan, L., YAN, S.: Moe++: Accelerating mixture-of-experts methods with zero-computation experts. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=t7P5BUKcYv

[22] Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024), https://kellerjordan.github.io/posts/muon/

[23] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020), https://arxiv.org/abs/2001.08361

[24] Kar, O.F., Tonioni, A., Poklukar, P., Kulshrestha, A., Zamir, A., Tombari, F.: Brave: Broadening the visual encoding of vision-language models. In: European Conference on Computer Vision. pp. 113–132. Springer (2024)

[25] Krell, M.M., Kosec, M., Perez, S.P., Fitzgibbon, A.: Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. arXiv preprint arXiv:2107.02027 (2021)

[26] Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension (2023), https://arxiv.org/abs/2307.16125

[27] Li, K., Jiang, Z., Cao, X., Wang, J., Yuchen, X., Meng, D., Wang, Z.: Describeearth: Describe anything for remote sensing images. arXiv preprint arXiv:2509.25654 (2025)

[28] Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Mini-gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814 (2024)

[29] Li, Z., Chen, G., Liu, S., Wang, S., VS, V., Ji, Y., Lan, S., Zhang, H., Zhao, Y., Radhakrishnan, S., Chang, N., Sapra, K., Deshmukh, A.S., Rintamaki, T., Le, M., Karmanov, I., Voegtle, L., Fischer, P., Huang, D.A., Roman, T., Lu, T., Alvarez, J.M., Catanzaro, B., Kautz, J., Tao, A., Liu, G., Yu, Z.: Eagle 2: Building post-training data strategies from scratch for frontier vision-language models (2025), https://arxiv.org/abs/2501.14818

[30] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/

[31] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

[32] Liu, S., Fan, L., Johns, E., Yu, Z., Xiao, C., Anandkumar, A.: Prismer: A vision-language model with multi-task experts (2024), https://arxiv.org/abs/2303.02506

[33] Man, Y., Huang, D.A., Liu, G., Sheng, S., Liu, S., Gui, L.Y., Kautz, J., Wang, Y.X., Yu, Z.: Argus: Vision-centric reasoning with grounded chain-of-thought. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14268–14280 (2025)

[34] Maniparambil, M., Akshulakov, R., Djilali, Y.A.D., Narayan, S., Singh, A., O'Connor, N.E.: Harnessing frozen unimodal encoders for flexible multimodal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)

[35] Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)

[36] OpenAI: Gpt-5 system card. Tech. rep., OpenAI (Aug 2025), https://cdn.openai.com/gpt-5-system-card.pdf

[37] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., et al.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774

[38] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193

[39] Pakov, T.: vit-face-expression (revision 78ed8d3) (2024). https://doi.org/10.57967/hf/2289, https://huggingface.co/trpakov/vit-face-expression

[40] Park, N., Kim, W., Heo, B., Kim, T., Yun, S.: What do self-supervised vision transformers learn? (2023), https://arxiv.org/abs/2305.00729

[41] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

[42] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

[43] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer (2017), https://arxiv.org/abs/1701.06538

[44] Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Huang, D.A., Yin, H., Sapra, K., Yacoob, Y., Shi, H., Catanzaro, B., Tao, A., Kautz, J., Yu, Z., Liu, G.: Eagle: Exploring the design space for multimodal llms with mixture of encoders. arXiv:2408.15998 (2024)

[45] Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Zhao, Y., Huang, D.A., Yin, H., Sapra, K., Yacoob, Y., Shi, H., Catanzaro, B., Tao, A., Kautz, J., Yu, Z., Liu, G.: Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=Y2RW9EVwhT

[46] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)

[47] Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)

[48] Team, K.: Kimi k2: Open agentic intelligence (2025), https://arxiv.org/abs/2507.20534

[49] Team, M.L.: Longcat-flash technical report (2025), https://arxiv.org/abs/2509.01322

[50] Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388

[51] Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)

[52] Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)

[53] Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025), https://arxiv.org/abs/2502.14786

[54] Wang, W., Gao, Z., Gu, L., Pu, H., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265

[55] Wu, A., Yang, Y., Luo, X., Yang, Y., Wang, C., Hu, L., Dai, X., Chen, D., Luo, C., Qiu, L., et al.: Llm2clip: Powerful language model unlock richer visual representation. In: NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice (2024)

[56] xAI: Grok-1.5v preview: Our first-generation multimodal model. https://x.ai/news/grok-1.5v (Apr 2024), accessed YYYY-MM-DD

[57] Xu, H., Xie, S., Tan, X., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying CLIP data. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=5BCFlnfE1g

[58] Yang, A., Yang, B., Hui, B., Zheng, B., et al.: Qwen2 technical report (2024), https://arxiv.org/abs/2407.10671

[59] Zhang, D., Long, T., Zhang, J., Chen, L., Ren, J., Zhang, Y.: Prompt-aware of frame sampling for efficient text-video retrieval (2025), https://arxiv.org/abs/2507.15491

[60] Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)

[61] Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8281–8291 (2024)

[62] Zhao, C., Xu, Y., Chen, Z., Gu, E., Zhang, K., Liu, X., Yang, J., Tai, Y.: From zero to detail: A progressive spectral decoupling paradigm for uhd image restoration with new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)

[63] Zhong, S., Gao, S., Huang, Z., Wen, W., Zitnik, M., Zhou, P.: Moextend: Tuning new experts for modality and task extension (2024), https://arxiv.org/abs/2408.03511

[64] Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems **35**, 7103–7114 (2022)

PolyVision: Conditional Visual Scaling Via Dynamic Expert Routing For Vision-Centric MLLMs

Supplementary Material

S1 VisPack Implementation Details

In this section, we provide a concise description of the **VisPack** mechanism. VisPack enables *variable-length expert execution* by grouping routed tokens into compact expert-specific sequences, while enforcing strict inter-sample isolation through block-diagonal attention masks.

S1.1 Expert-wise Token Collection

Given the routing mask $\mathcal{M} \in \{0,1\}^{E\times B\times S}$ and CLS routes $\{g_b\}$, all tokens assigned to expert i are gathered into an index set

$$\mathcal{I}_i = \{(b,s) \mid \mathcal{M}[i,b,s] = 1\} \cup \{(b,-1) \mid g_b = i\},$$

where $s = -1$ denotes CLS tokens. Trunk experts collect both CLS and patch tokens, while non-trunk experts collect only patch tokens.

S1.2 Variable-length Packing

To minimize padding and redundant computation, each index set $\mathcal{I}_i$ is greedily partitioned into groups $\{G_{i,g}\}$ that satisfy

$$|G_{i,g}| \le T_{\max}.$$

Each group is then packed into a contiguous token sequence

$$P_{i,g} = [\, x_{b,s} \text{ or } c_b \,]_{(b,s)\in G_{i,g}},$$

with an associated scatter map $\Gamma_{i,g}$ that records the original sample/token positions.

S1.3 Block-diagonal Attention Mask

Since groups may contain tokens from different samples, VisPack constructs a block-diagonal attention mask:

$$A_{i,g}[u,v] = \mathbf{1}[\, b(\Gamma_{i,g}[u]) = b(\Gamma_{i,g}[v]) \,]. \tag{S1}$$

This ensures no cross-sample information flow during the expert's self-attention. Under the condition of `CLIP+DINO+Zero Expert`, the visualization of the attention mask is shown in Figure S1.

S1.4 Pseudocode

For completeness, we replicate Algorithm S1, which summarizes the entire workflow.

S2 Training Details

Table S1 summarizes the four-stage training pipeline of `PolyVision`. Below we provide a concise description of each stage, including objectives, datasets, optimization settings, and trainable components.

S2.1 Stage 1: Visual Pre-alignment

This stage focuses on aligning heterogeneous visual experts to a unified embedding space before interacting with the language model. We train on CC3M and LAION, totaling approximately 5M image–caption pairs. Only the vision experts are updated. A differential learning-rate scheme is applied: newly introduced modules are optimized with a higher learning rate (2×10^{-3}), while pretrained components are updated more conservatively using a lower learning rate (2×10^{-5}). We

Algorithm S1 VisPack: Variable-length Token Packing

Require: Patch tokens $\{x_{b,s}\}$, CLS tokens $\{c_b\}$, routing mask $\mathcal{M}$, patch routing weights $\{w_{b,s,i}\}$, CLS routes $\{g_b\}$, token limit $T_{\max}$
Ensure: Packed groups $\{P_{e,g}\}$, masks $\{A_{e,g}\}$, scatter maps $\{\Gamma_{e,g}\}$

for $i = 1..E$ **do**
  $\mathcal{I}_i \leftarrow \{(b,s) \mid \mathcal{M}[i,b,s] = 1\} \cup \{(b,-1) \mid g_b = i\}$
end for
for $i = 1..E$ **do**
  Split $\mathcal{I}_i$ into $\{G_{i,g}\}$ with $|G_{i,g}| \le T_{\max}$
  for $G_{i,g}$ **do**
    $P_{i,g} \leftarrow [x_{b,s} \text{ or } c_b]_{(b,s)\in G_{i,g}}$
    $\Gamma_{i,g} \leftarrow \{(b,s)\}_{(b,s)\in G_{i,g}}$
  end for
end for
for each (i,g) **do**
  $A_{i,g}[u,v] = \mathbf{1}[\, b(\Gamma_{i,g}[u]) = b(\Gamma_{i,g}[v])\,]$
end for
for each (i,g) **do**
  $Y_{i,g} \leftarrow \mathrm{Expert}_i(P_{i,g}; \text{attention_mask} = A_{i,g})$
  Scatter-add $w_{b,s,i} Y_{i,g}[u]$ to $(b,s) = \Gamma_{i,g}[u]$ for patch tokens; scatter CLS outputs once.
end for

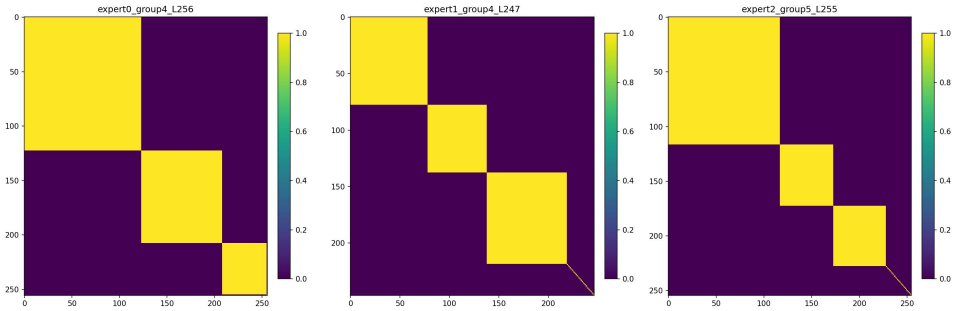


Figure S1: **Block-diagonal attention mask visualization.** In the figure, we can clearly observe that although only a simple greedy strategy is used, the padding in VisPack only a small portion.

adopt a global batch size of 18K and employ standard image–text contrastive and reconstruction objectives.

S2.2 Stage 2: Projector Warmup

To stabilize subsequent multimodal alignment, we warm up the vision–language projector using the LLaVA-en-zh-300k dataset. Only the adapter (projector) parameters are trainable in this stage, while both the vision experts and the language model remain frozen. We use a learning rate of 1×10^{-4} and a batch size of 256, training on roughly 300K image–instruction pairs. This warmup ensures that the projector reliably maps visual embeddings into the language model space *without perturbing the pretrained representations of either modality*, thereby preventing instability or representation drift in downstream alignment.

S2.3 Stage 3: Vision–Language Alignment

This stage aligns `PolyVision` vision experts and projector jointly with the language model using the Cambrian-Alignment dataset (2.5M samples). Both the visual experts and the adapter are trainable, while the language model remains frozen. We adopt a learning rate of 1×10^{-5} and a global batch size of 4096. Notably, the Cambrian-Alignment corpus primarily consists of short, concise instruction–response pairs, which makes it well suited for stabilizing early-stage multimodal grounding. This stage therefore establishes the core visual–language alignment necessary for downstream capabilities.

Table S1: **Overview of `PolyVision` training stages.** See supplementary material for detailed configurations.

Stages	Stage 1	Stage 2	Stage 3	Stage 4
Purpose	Visual Pre-alignment	Projector Warmup	Vision-Language Align	Multimodal Pre-training
Dataset	CC3M, LAION	LLaVA-300K	Cambrian-Align	Cambrian-7M
Learning Rate	2e-3, 2e-5	1e-4	1e-5	2.5e-5
Batch Size	18K	256	4096	128
Trainable	Vision Experts	Adapter	Experts & Adapter	All

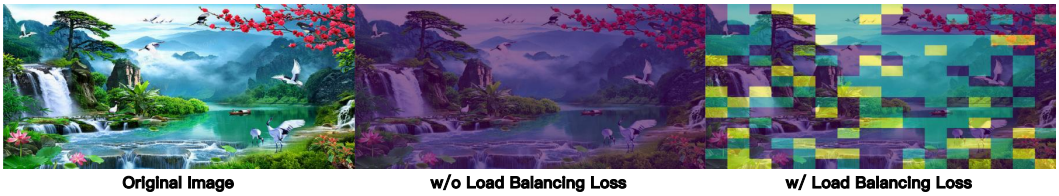


Figure S2: **The visualization of attentive router selection.** When not using the load balancing loss, the model tends to take shortcuts and collapses to a single expert, typically a language-aligned model such as CLIP.

S2.4 Stage 4: Multimodal Pre-training

The final stage performs broad multimodal pre-training using predominantly general-purpose and diverse-task datasets. We train on Cambrian-7M, covering a wide range of multimodal understanding and reasoning scenarios. All model components, including visual experts, adapters, and the language model, are jointly optimized. We use a learning rate of 2.5×10^{-5} and a batch size of 128 over a total of 7M training samples. This stage improves general multimodal reasoning, robustness, and instruction-following ability across heterogeneous tasks.

S3 Discussion

S3.1 Scaling Behavior

By independently controlling the number of experts, heterogeneity, dataset size, and text encoder size during stage 1, we obtained the results shown in Figure 6 (b). For a fair comparison, we uniformly compress the channel dimension to 1024 and use the CLIP Loss with continuous targets on the ImageNet1K [42] val set as the analysis target.

Diversity of experts matters. In this experiment, we adopt two settings: (1) progressively adding heterogeneous experts, and (2) progressively adding homogeneous experts. We observe that increasing heterogeneous experts yields consistent and stable improvements, whereas adding more homogeneous experts leads to marginal gains that quickly saturate. This indicates that performance gains stem primarily from representational diversity rather than mere capacity expansion, as heterogeneous experts provide complementary features, whereas homogeneous experts yield rapidly saturating returns due to representational redundancy.

Alignment data scaling is bounded by model capacity. We further vary the size of the alignment dataset from 1M to 5M and 15M samples. Increasing the data scale brings noticeable performance gains in the early stages; however, the benefit diminishes as the model capacity becomes a limiting factor, resulting in slower convergence and negligible improvement at larger scales.

Vision scaling dominates. Finally, we scale the text embedding model from 0.6B to 4B and 8B parameters. Enlarging the text model yields clear and measurable gains in stage 1. Nevertheless, the improvement per parameter increase remains noticeably smaller than that achieved by expanding the visual components, suggesting that the predominant performance bottleneck in multimodal learning still resides on the visual side.

S3.2 Pre-alignment with LLM2VEC

Many recent studies have highlighted the broad benefits of aligning LLM2VEC models with CLIP [55], such as improved compatibility with longer texts and the emergence of multilingual capabilities even when trained solely on monolingual datasets. These properties enhance the model's adaptability to downstream tasks. In our small-scale experiments, we observed that employing the

Qwen3-Embedding model as the text tower yielded better pretraining performance compared to the CLIP-Text tower. We hypothesize that this improvement arises because our dataset contains a mixture of long and short text samples, allowing embedding models to better capture complex distributions. We further compared the effects of different prompt strategies applied to the embedding model and found that using **summarization-style prompts** led to the best performance. Specifically, the prompt we used during the alignment process was: *Provide a description of an image that captures its most prominent features for recognition, please summarize its meaning in one word.*

A likely explanation is that summarization-style prompts encourage the embedding model to focus on high-level semantics rather than surface wording, yielding more stable and modality-consistent representations. This reduces stylistic variance across mixed long–short texts and produces embeddings that align more naturally with visual features, leading to improved pretraining stability and performance.

S3.3 Optimizer

Throughout the entire training process, we observed a high degree of heterogeneity introduced by our model, which motivated us to adopt the newly proposed optimizer `Muon` [22]. Building upon the adaptive strategy of Kimi-K2 [48], we made several modifications: specifically, we retained the AdamW optimizer for both the embedding and head components, while applying a differential learning rate scheme to the pretraining modules, scaling the base learning rate by a factor of 100 and adjusting it according to matrix size as adjusted_lr $= \text{lr} \times 0.2 \times \sqrt{\max(A, B)}$. Although previous work has reported that `Muon` may yield suboptimal results when used as an SFT optimizer [48], our small-scale experiments indicate that, within our model, `Muon` still outperforms AdamW. Therefore, we retain this optimization strategy in our training pipeline.

Notably, in the first stage of training, we apply a weight decay of 0.02 and use a linear warmup for the first 1 epoch to gradually reach the target learning rate. After warmup, the schedule follows a cosine decay strategy throughout the remainder of training.

S3.4 Mixture of Experts

In our experiments, we found that many conclusions drawn from traditional MoE frameworks still hold in our model. Moreover, several MoE-based enhancements, such as the *zero experts* mechanism proposed in Long-cat [49] and MoE++ [21], can be directly applied to our approach as a free-lunch improvement, bringing notable efficiency gains. In the main text, we enable zero experts by default to further enhance efficiency. This naturally raises an intriguing question: can our method be extended to broader scenarios and combined with existing MoE techniques? The evidence so far suggests that such integration could lead to even greater flexibility and potential.

S4 Ablation Study

S4.1 Ablation on Load Balancing Loss

We next remove the load balancing regularization term that enforces a uniform distribution of token assignments across experts. This modification causes the routing weights to collapse toward a dominant expert, as visualized in Figure S2. Quantitatively, the model's performance degenerates to that of a single CLIP-based encoder (Table S2), dropping from 23.0 to 13.3 on MMVP Average. The underlying reason is that without load balancing, the router optimizes purely for local performance and converges to a trivial solution where most tokens are routed to the strongest expert, effectively disabling the multi-expert specialization. This imbalance severely reduces diversity in visual feature representations and limits the model's capacity to capture heterogeneous visual semantics. During the experiment, we set the load balancing loss coefficient to 0.01.

Table S2: **Ablation on the load-balancing loss.**

Model	CLIP	Ours	w/o aux
MMVP	19.3	**23.0**	13.3

S4.2 Number of Activated Experts

We set the number of activated experts to two after examining both intuitive and empirical factors. When only a single expert is activated, the model is forced to map all visual patterns, semantic regions, high-frequency structures, textures, and fine-grained details into a single representational space. Intuitively, many visual regions exhibit a mixture of semantic and low-level cues, and relying on a single expert prevents the model from capturing these heterogeneous properties effectively. This typically leads to oversimplified representations and a clear drop in performance. Conversely, activating more than two experts increases computation and memory consumption while offering diminishing returns, and may even reduce specialization by diffusing token assignments across too many experts. Activating two experts provides a balanced configuration: it allows complementary experts to jointly model diverse visual characteristics while preserving computational efficiency and maintaining strong expert specialization.

S4.3 Different Alignment Losses

CLIP Contrastive Loss. We adopt the standard global contrastive formulation widely used in CLIP-style pre-training [41]. Given a batch of B image–text pairs, the fused visual [CLS] feature $\bar{v}_i$ and the (frozen) CLIP text embedding $\bar{t}_i$ form aligned pairs. The objective maximizes the similarity between corresponding pairs while suppressing similarity to other samples in the batch:

$$L_{\text{CLIP}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp\big(\phi(\bar{v}_i,\bar{t}_i)/\tau\big)}{\sum_{j=1}^{B}\exp\big(\phi(\bar{v}_i,\bar{t}_j)/\tau\big)} + \log\frac{\exp\big(\phi(\bar{t}_i,\bar{v}_i)/\tau\big)}{\sum_{j=1}^{B}\exp\big(\phi(\bar{t}_i,\bar{v}_j)/\tau\big)}\right] \tag{S2}$$

where $\phi(x,y) = \frac{x}{\|x\|_2}\cdot\frac{y}{\|y\|_2}$ is cosine similarity and τ is a temperature parameter. This loss converges quickly in early training, but the static linguistic prior of the frozen text tower causes alignment to saturate rapidly, forming a performance bottleneck for deeper multimodal learning.

SigLIP Loss. SigLIP replaces the softmax-normalized InfoNCE with independent sigmoid-based binary classification over all positive and negative pairs. Formally, let $z_{ij} = \alpha\,\phi(v_i,t_j) + \beta$, where $\alpha > 0$ and β are the learnable logit scale and bias. For each image–text pair (v_i,t_i):

$$L_{\text{SigLIP}} = -\frac{1}{B}\sum_{i=1}^{B}\Big[\log\sigma(z_{ii}) + \sum_{j\neq i}\log\big(1-\sigma(z_{ij})\big)\Big] \tag{S3}$$

and symmetrically for t_i as the anchor. Although more flexible, SigLIP introduces higher gradient variance during multimodal feature evolution, leading to slower convergence and weaker final results compared to our LLM2VEC-based loss.

Patch-Level Contrastive Loss. For patch-level alignment [5], we closely follow the formulation used in SPARC, which computes fine-grained matching between text token embeddings and patch-level visual features. For each text token t_{ij} in sample i, patch–token alignment weights are defined as:

$$a_{jk} = \frac{\tilde{s}_{jk}}{\sum_{r=1}^{R}\tilde{s}_{jr}}, \qquad c_l = \sum_{r=1}^{R} a_{lr}v_r \tag{S4}$$

where $\tilde{s}_{jk}$ is a sparsified similarity score between token j and patch k, and R is the number of non-zero weights. A language-grouped visual embedding is constructed as:

$$L_{\text{patch}} = -\frac{1}{2B}\sum_{i=1}^{B}\frac{1}{L_i}\sum_{j=1}^{L_i}\left[\log\frac{\exp\big(\phi(c_{ij},t_{ij})/\tau\big)}{\sum_{k=1}^{L_i}\exp\big(\phi(c_{ij},t_{ik})/\tau\big)} + \log\frac{\exp\big(\phi(t_{ij},c_{ij})/\tau\big)}{\sum_{k=1}^{L_i}\exp\big(\phi(t_{ij},c_{ik})/\tau\big)}\right] \tag{S5}$$

Despite its conceptual appeal, this loss diverges consistently in our training, producing `NaN` values early. We attribute this to (1) the granularity mismatch between local patch features and globally contextualized text semantics, and (2) severe gradient imbalance caused by patch-level competition within each image–text pair.

LLM2VEC Contrastive Loss (Ours). Our adopted loss uses contextualized last-token embeddings from a large language model (LLM2VEC) and aligns them with fused visual [CLS] features. For each sample, let h_i denote the LLM text embedding and f_i the fused visual representation. The alignment is performed via:

$$L_{\text{LLM2VEC}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp\big(\phi(f_i,h_i)/\tau\big)}{\sum_{j=1}^{B}\exp\big(\phi(f_i,h_j)/\tau\big)} + \log\frac{\exp\big(\phi(h_i,f_i)/\tau\big)}{\sum_{j=1}^{B}\exp\big(\phi(h_i,f_j)/\tau\big)}\right] \tag{S6}$$

Unlike CLIP and SigLIP, the LLM text embedding evolves with contextual content, offering adaptive semantic priors. This avoids saturation and enables stronger cross-modal grounding, achieving the best stability and final performance.

SigLIP converges more slowly due to noisy independent binary gradients; CLIP converges fast but saturates early because its frozen text tower limits semantic richness; Patch-Level Contrastive Loss consistently diverges to `NaN`, likely due to local–global mismatch and unstable token-level gradients; our LLM2VEC-based loss provides stable training, avoids early plateaus, and offers richer contextual alignment, making it the most effective strategy.

S5 Broader Impact and Potential Misuse

In this paper, we introduced `PolyVision`, a novel multimodal large language model with strong cross-modal understanding. Our experiments show that `PolyVision` achieves state-of-the-art performance on several benchmarks, validating the effectiveness of our architecture and training strategy. On the positive side, stronger vision-centric multimodal models can support applications such as assistive reading, document understanding, chart analysis, and other settings that benefit from fine-grained visual perception. Despite its strong results, we acknowledge inherent limitations and risks. A key issue is **hallucination**, where the model may produce factually incorrect or context-inconsistent content. Such errors may arise from dataset biases, imperfections in visual encoding, or modality misalignment during training. In addition, the model's powerful generative ability introduces risks of **misuse**, including misinformation, amplification of societal biases, and potential privacy or copyright concerns. To mitigate these risks, we primarily rely on **curated and openly licensed data** for training and adopt **responsible-use policies** that discourage harmful applications. We will continue to refine data quality, alignment mechanisms, and evaluation practices to enhance safety and reliability.

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: The abstract and introduction state the paper's motivation, method, and empirical scope, and these claims are supported by the Method, Experiments, and Conclusion sections.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The paper discusses limitations and risks, including the current lack of RoPE-based multi-expert integration and potential failure modes such as hallucination and misuse; see the Conclusion and Supplementary Sec. *Broader Impact and Potential Misuse*.

   Guidelines:

   - The answer [N/A] 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: [N/A]

Justification: The paper does not present formal theorems, lemmas, or proofs; its technical contribution is architectural and empirical.

Guidelines:

- The answer [N/A] 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: The manuscript describes the architecture, routing and packing procedures, expert combinations, datasets, training stages, optimizer choices, and key ablations needed to reproduce the main claims; see Method, Experiments, and Supplementary Secs. Training Details, Optimizer, and Ablation Study.

Guidelines:

- The answer [N/A] 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: An anonymized code package is provided in the supplementary material for review, and the experiments use cited public datasets and benchmarks; we plan to release the code and model weights publicly after the review process.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the benchmark suite, expert combinations, stage-wise datasets, learning rates, batch sizes, trainable modules, optimizer choice, scheduling details, and several ablation settings; see Experimental Setup and Supplementary Secs. Training Details, Optimizer, and Ablation Study.

   Guidelines:

   - The answer [N/A] 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: The paper reports point estimates only and does not include repeated-run statistics, confidence intervals, significance tests, or error bars for the main experimental results.

Guidelines:

- The answer [N/A] 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: [No]

   Justification: The paper compares efficiency and memory behavior, but it does not report the exact compute worker type, memory capacity, or wall-clock time required for each experiment.

   Guidelines:

   - The answer [N/A] 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: To the best of our knowledge, the research complies with the NeurIPS Code of Ethics; it does not involve human-subject experimentation and it discusses broader impacts, potential misuse, and mitigation in Supplementary Sec. *Broader Impact and Potential Misuse*.

   Guidelines:

   - The answer [N/A] 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: [Yes]

    Justification: Supplementary Sec. *Broader Impact and Potential Misuse* discusses both positive impacts, such as assistive reading and visual understanding applications, and negative impacts, such as hallucination, misuse, privacy, and copyright concerns.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Supplementary Sec. *Broader Impact and Potential Misuse* states that training relies on curated, openly licensed data and that release will follow responsible-use policies intended to discourage harmful applications.

    Guidelines:

    - The answer [N/A] 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: [No]

    Justification: The paper credits prior datasets, models, and benchmarks through citations, but it does not explicitly enumerate the licenses and terms of use for all external assets used in the work.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The submission does not provide a new dataset, code package, or model artifact alongside the paper; it only states that code and weights will be released in the future.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not report crowdsourcing experiments or other studies involving human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research that would require IRB review.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are an explicit part of the method: the model is built on Qwen3 or LLaMA backbones, and the visual pre-alignment stage uses Qwen3-Embedding as an LLM2VEC text tower; see the Abstract, Introduction, Visual Pre-Alignment, and Supplementary Sec. Pre-alignment with LLM2VEC.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.