Boosting Multimodal LLMs via Visual Token Supervision

Zhipeng Bao[1,2] Miao Liu[1] Ankit Ramchandani[1] Mengjiao Wang[1] Felix Juefei-Xu[1]
Xide Xia[1] Tong Xiao[1] Yu-Xiong Wang[3] Martial Hebert[2] Ning Zhang[1] Xiaofang Wang[1]
[1]GenAI, Meta [2]Carnegie Mellon University [3]University of Illinois Urbana-Champaign

Abstract

*Multimodal large language models (MLLMs) have shown impressive performance on tasks requiring integrated visual and textual understanding. A key factor in their success is the model's ability to accurately recognize and understand visual elements. While recent advancements focus on enhancing vision encoders to produce richer visual tokens, an often overlooked aspect is how effectively the underlying language model can further process these visual tokens. Through a vision-centric analysis, we find that the intermediate visual representations of MLLMs perform poorly on semantic and geometric understanding tasks, even worse than their standalone vision encoders. More importantly, our analysis reveals that the quality of visual tokens of MLLMs begins to degrade even before being processed by the language model, indicating inherent flaws in the current MLLM designs. To address this, we introduce a self-distillation approach to refine the visual tokens of MLLMs through a reverse multimodal projector, enhancing alignment with original visual features. Extensive evaluations confirm that our method generalizes across diverse MLLM architectures, scales effectively, and consistently improves MLLMs' performance on perception-oriented benchmarks (*e.g.*, SEED, Real-WorldQA, CV-Bench) while maintaining overall performance on general-purpose benchmarks (*e.g.*, *MMMU, ChartQA, MMB).*

1. Introduction

Multimodal large language models (MLLMs) [26, 31, 46] have demonstrated impressive performance on tasks that require integrated understanding and reasoning across visual and textual modalities. Their success largely hinges on a robust visual recognition capability – accurately identifying objects, scenes, and other key elements in images – to generate relevant and precise responses. Recent work has largely focused on enhancing this capability by employing more powerful vision encoders that produce richer visual tokens for the underlying language model to process[26, 46]. Yet, a critical question remains largely unexplored: *How effectively does the language model refine and utilize these visual tokens?*

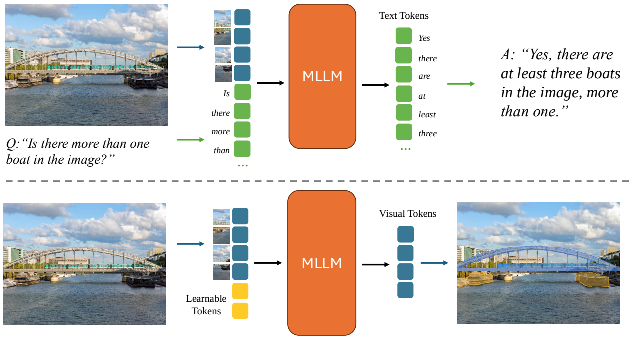
Text Tokens
Yes
there
are
at
least
three
...
MLLM
Is
there
more
than
...
Q: "Is there more than one boat in the image?"
A: "Yes, there are at least three boats in the image, more than one."
MLLM
Visual Tokens
Learnable Tokens

Figure 1. **Top:** Conventional VQA-based MLLM evaluation, which assesses whether the generated text tokens match the correct answer. **Bottom:** Our proposed vision-centric evaluation, which measures the quality of visual tokens on downstream vision tasks.

To address this critical question, we conduct a targeted analysis (detailed in Section 2) that deviates from conventional evaluation methods. Rather than relying on visual question answering (VQA) to evaluate model performance, our approach directly assesses the quality of visual tokens within MLLMs. Inspired by established practices in computer vision [14, 15], we extract the visual tokens from multiple layers of the model, treating them as intermediate representations, and attach lightweight probing heads to evaluate performance on downstream tasks. Specifically, we assess these representations on semantic segmentation using ADE20K [65] and depth estimation on CityScapes [8]. This decoupled evaluation method isolates visual recognition from language generation, revealing that intermediate representations not only underperform compared to the original vision encoder, but also degrade in quality **even before being processed by the language model**. These unexpected findings expose inherent flaws in current MLLM architectures and underscore the necessity for methods that enhance the quality of visual tokens to ultimately improve the overall performance of these models.

Motivated by this observation, we propose a novel self-distillation loss to directly supervise and enhance the quality of visual tokens. In our approach, as illustrated in Fig-

Visual Tokens	CLIP	MM Projector (L0)	L10	L20	L30	L40
mIoU@ADE20K (↑)	**32.02**	25.51	28.32	29.03	28.40	27.04
mErr@CityScape (↓)	**4.92**	6.43	5.95	5.62	5.58	5.42

Table 1. Performance of visual tokens on downstream vision tasks. Visual tokens from MLLMs exhibit weak semantic and geometric understanding, even performing worse than the original vision encoder. This degradation ultimately affects the accuracy of language responses that rely on these tokens for image understanding.

ure 3, we augment standard MLLM architectures – which typically comprise a vision encoder, a multimodal projector, and a language model [31, 46, 54] – with an additional reverse multimodal projector. This module maps the visual tokens back into the original visual feature space, and a cosine similarity loss is applied to enforce alignment between the recovered tokens and the initial visual features. Such a design encourages the visual kens to capture richer information, thereby boosting the MLLM's visual understanding and improving the accuracy of its language outputs on vision-language tasks. Through extensive experimental evaluations, we validate that:

- Our approach generalizes across diverse MLLM architectures, scales effectively, and consistently improves performance across all configurations.
- Adding the self-distillation loss consistently improves performance on perception-oriented benchmarks (*e.g.*, SEED, Real-WorldQA, CV-Bench) and achieves state-of-the-art results within mid-size models, without compromising performance on general-purpose benchmarks (*e.g.*, MMMU, ChartQA, MMB).
- Leveraging additional visual expert models as sources for distillation further enhances performance on various MLLM benchmarks.

By directly supervising the visual tokens, our work not only provides a more transparent evaluation of MLLM visual representations but also addresses a critical shortcoming in current architectures. We believe this approach helps foster incremental improvements in vision-language understanding for future MLLMs.

2. Analysis of Visual Tokens in MLLMs

We introduce a vision-centric analysis framework to systematically examine the visual tokens in MLLMs as illustrated in Figure 1. Unlike conventional MLLM evaluation methods that rely on VQA-style benchmarks, our approach directly assesses MLLMs' visual recognition capability by evaluating their intermediate representations (visual tokens) on downstream vision tasks.

2.1. Evaluation Setup

We conduct the analysis with a pre-trained MLLM [32], which employs CLIP [41] as the vision encoder and

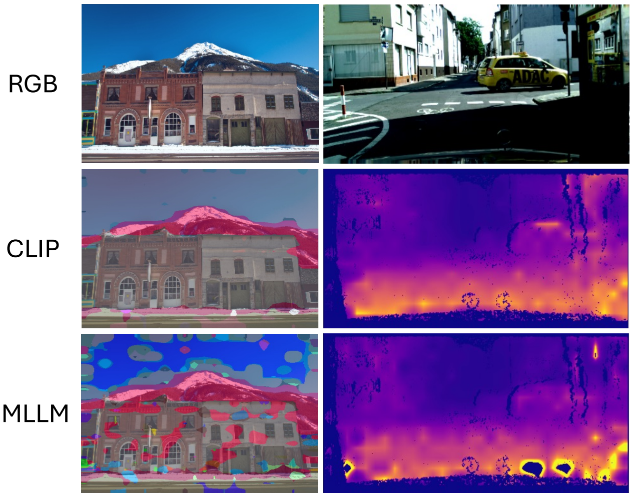

Figure 2. Qualitative comparisons between the intermediate visual representations of a MLLM and its visual encoder, CLIP [41]. Surprisingly, we find that these visual representations perform poorly on semantic segmentation and depth estimation – even underperforming its vision encoder alone.

Vicuna-7B [64] as the base language model. We evaluate two representative vision tasks: semantic segmentation on ADE20K [65] and depth estimation on CityScapes [8]. Inspired by DINOv2 [38], we attach a lightweight two-layer MLP head to the extracted visual tokens to predict dense segmentation and depth maps, which are then interpolated to the original resolution for evaluation.

Conceptually, the visual tokens in MLLMs should be well-suited for these vision tasks, as they could be seen as analogous to instructing an MLLM to describe each pixel in the image sequentially. However, since pre-trained MLLMs are not explicitly trained for such structured pixel-wise tasks, we introduce two learnable task tokens. These tokens act as soft prompts, guiding the model toward effective feature extraction for vision tasks [22]. Our experiments confirm that adding learnable task tokens improves performance, but further increasing their numbers does not provide further gains in vision tasks. During our evaluation, only the task head and the task tokens are optimized, while all other model weights remain frozen.

2.2. Results and Analysis

We first compare the final-layer visual tokens from the MLLM (output of the last self-attention layer) to those extracted directly from its vision encoder, CLIP. Qualitative and quantitative results are provided in Figure 2 and Table 1 (second and final columns, respectively). We then extend our analysis to visual tokens from different layers of the MLLM (as shown in other columns of Table 1), including those from the multimodal (MM) projector [31], which serves as the input to the large language model (LLM) and can be viewed as layer 0 representations. Our key observations include:

- Visual tokens within MLLMs underperform their own vision encoder on downstream vision tasks. Qualitatively, MLLMs exhibit limited semantic comprehension, often producing imprecise object boundaries.
- The tokens from the MM projector show the weakest performance on vision tasks, suggesting that **visual information loss occurs before the LLM even begins processing the tokens**. This suggests that the LLM receives an already degraded representation of the input image, limiting its ability to perform fine-grained visual reasoning.
- Visual token quality improves as they propagate deeper into the LLM, implying that the model can partially restore lost information. However, this recovery remains incomplete, meaning that critical visual details may still be missing from the final representation.

These findings suggest that existing MLLM architectures do not fully leverage the information extracted by their vision encoders. While MLLMs are primarily designed for language-based reasoning, high-quality visual representations are equally crucial, particularly for vision-centric tasks. Richer visual tokens enable MLLMs to capture fine-grained spatial and compositional details, which directly impact their ability to generate accurate language outputs.

Motivated by these insights, we propose a method to preserve and enhance visual token quality, ensuring that essential visual information is retained before being processed by the LLM. By mitigating information loss at the multimodal projector stage, our approach strengthens the model's visual-language reasoning capabilities and improves its performance on vision-centric tasks.

In the supplementary material, we extend this analysis to stronger MLLMs, demonstrating the consistency and robustness of the findings drawn in this section.

3. Method

In this section, we begin with an overview of the foundational LLaVA-like architecture [31] used for MLLMs. We then present the details of our proposed self-distillation approach, as illustrated in Figure 3.

3.1. Preliminaries

The goal of an MLLM is to generate a textual response y based on multi-modal inputs, typically an image-question pair (x, c) [31]:

$$y = f_{\text{MLLM}}(x, c). \tag{1}$$

As shown in Figure 3 (Loop marked in black), a typical MLLM architecture consists of the following main components: a vision encoder to extract image features, a multimodal projector to map visual features into the language space, a text tokenizer to create text token embeddings, and a LLM to integrate multimodal features and produce the final response. To process visual information, the MLLM first extracts features from the input image x and projects these features into the language space, which are then passed to the LLM as input visual tokens:

$$x_{\text{input}} = \theta_{\text{mm}}(\mathcal{E}_{\text{img}}(x)), \tag{2}$$

where $\mathcal{E}$ is the vision encoder, typically a frozen CLIP-based model [41, 61], and θ_{mm} represents the multimodal projector, often implemented as an MLP or Perceiver [18].

Next, the MLLM takes both the text and visual tokens as input, processes them through several self-attention layers [49], and generates output tokens iteratively. Finally, the tokenizer is used to decode the output tokens into the final textual response:

$$y = \mathcal{D}(\psi_{\text{LLM}}(x_{\text{input}}, \mathcal{E}_{\text{t}}(c))), \tag{3}$$

where ψ_{LLM} denotes the LLM network, $\mathcal{E}_t$ represents the text tokenizer encoder and $\mathcal{D}$ is the tokenizer decoder.

Following the approach in LLaVA [31], MLLMs are typically trained using a two-stage process. In the first stage, called the *pre-training* stage, only the multimodal projector is fine-tuned to align visual representations with the language space. In the second stage, known as *instruction tuning*, all components except for the vision encoder and tokenizer are further fine-tuned. We adopt this training schedule in our experiments. For both stages, the MLLM is trained with a next-token prediction objective:

$$\mathcal{L}_{\text{LLM}}(\theta_{\text{mm}}, \psi_{\text{LLM}}) = -\sum_{i=1}^{N} \log P(y_i | y_{:i-1}, x_{\text{input}}, \mathcal{E}_{\text{t}}(c)), \tag{4}$$

where y_i denotes the i^{th} token in the target response y, N is the total number of tokens in the response, and $y_{:i-1}$ are the preceding tokens before y_i. This objective calculates the negative log-likelihood of the correct next token y_i conditioned on the previous text tokens and multimodal input. Note that this objective is only applied to θ_{mm} during the pre-training stage.

3.2. Visual Knowledge Distillation for MLLMs

To enhance the visual representations within MLLMs, we propose a self-distillation objective designed to preserve the richness of visual token information after processing by the LLM. Specifically, we introduce a reverse multimodal projector, applied to the output of the n^{th} self-attention layer, to map the visual representations back to the original visual feature space:

$$x_{\text{token}} = \theta_{\text{rmm}}\left(\psi^{(n)}_{\text{LLM}}(x_{\text{input}}, \mathcal{E}_{\text{t}}(c))\right), \tag{5}$$

where θ_{rmm} denotes the reverse multimodal projector, which mirrors the architecture of θ_{mm} except for a difference in

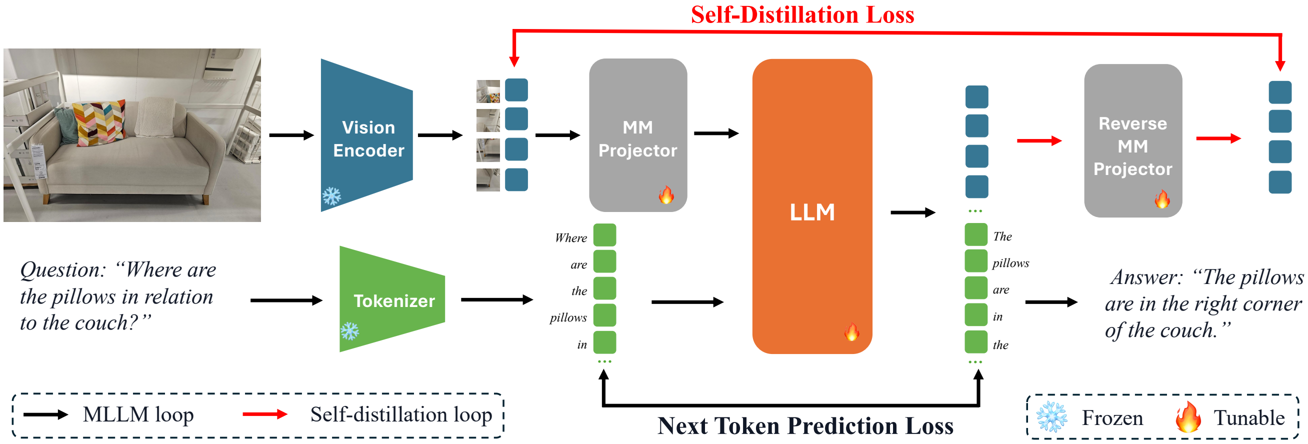

Figure 3. Architecture of our proposed self-distillation algorithm. **Loop marked in black:** a general architecture of MLLM; **Loop marked in red:** the proposed self-distillation objective that preserves the information of visual tokens. Our method enhances the vision-language reasoning capacity of MLLMs by providing a more informative and robust visual token representation.

hidden dimensions. Here, $\psi_{\text{LLM}}^{(n)}$ represents the intermediate visual representation after the n^{th} self-attention layer in the LLM. We then compute the cosine similarity between this projected visual representation and the initial representation extracted by the vision encoder. This similarity serves as the self-distillation loss, used to update the weights of the multimodal and reverse multimodal projectors:

$$\mathcal{L}_{\text{distill}}(\theta_{\text{mm}}, \theta_{\text{rmm}}) = 1 - \frac{x_{\text{token}} \cdot \mathcal{E}_{\text{img}}(x)}{\|x_{\text{token}}\| \cdot \|\mathcal{E}_{\text{img}}(x)\|}. \quad (6)$$

This self-distillation loss encourages the multimodal projector and LLM to retain critical information from the visual data. By incorporating the reverse multimodal projector and using cosine similarity as the objective, we also prevent the model from converging to trivial solutions.

Notably, this method does not require using visual features from the MLLM's native vision encoder alone; it is compatible with external visual expert models or a mixture of experts (MoE). In the case of M visual expert models, the self-distillation loss can be adapted as follows:

$$\mathcal{L}_{\text{distill}}(\theta_{\text{mm}}, \theta_{\text{rmm}}) = \sum_{i=1}^{M} \left(1 - \frac{x_{\text{token}} \cdot \mathcal{E}_i(x)}{\|x_{\text{token}}\| \cdot \|\mathcal{E}_i(x)\|}\right), \quad (7)$$

where $\mathcal{E}_i$ represents the i^{th} visual expert model, each of which may include its own vision encoder. We show the broader application of our approach with external visual expert models as sources of distillations in Section 4.4.

The final training objective, incorporating the proposed self-distillation loss, is:

$$\mathcal{L} = \mathcal{L}_{\text{LLM}} + \alpha \mathcal{L}_{distill}, \quad (8)$$

where α is a balancing factor between the two objectives.

4. Experimental Evaluations

4.1. Setup

Baseline model design. Our model builds upon pretrained vision encoders and LLMs. We employ SigLIP [61] as the vision encoder and QWen2-7B [55] as the LLM backbone. We denote this primary model as "Baseline" for our evaluations. During our method explorations, we also consider another variant of MLLM following the design of Liu et al. [32] using CLIP [41] as the vision encoder and Vicuna-7B [64] as the LLM.

Training data. Our full training dataset consists of 2M image-text pairs for pretraining and 6M samples for instruction tuning. The images are sourced from public domain, and the text responses are generated by a licensed MLLM. To assess the optimal integration of our self-distillation loss into MLLMs, we further consider a reduced version of the full dataset containing approximately 600K samples for pretraining and 700K samples for instruction tuning. This reduced dataset is also used for our ablation study.

Benchmarks. We primarily evaluate on perception-oriented benchmarks, including SEED [24], Real-WorldQA [52], CV-Bench^{2D}[46], MMVP[47], MME [10], and GQA [16]. For MME, we focus exclusively on perception tasks and denote this subset as MME$^{\text{p}}$.

To broaden our analysis, we also compare against state-of-the-art models on non-perception-oriented benchmarks, including MMB [35] and LLaVA-Bench [31] for general-purpose evaluation, MMMU [60] and AI2D [19] for assessing model knowledge, and TextVQA [44] and ChartVQA [36] for OCR understanding.

Further details on baseline configurations, evaluation benchmarks, compared models, and implementation are provided in the supplementary materials.

Stage1	Stage2	α	SEED	Real-WorldQA	CV-Bench^{2D}	MMVP	MME^P	GQA
✗	✗	-	64.4	58.2	60.4	37.4	1535.8	62.6
✓	✗	0.1	65.8	58.9	61.2	38.0	1574.6	62.8
✓	✗	1.0	64.7	58.3	60.8	37.4	1545.7	63.1
✓	✗	5.0	60.1	56.1	58.7	37.0	1460.7	60.8
✓	✓	0.1	**66.2**	**59.1**	61.6	**38.6**	**1578.0**	**63.5**
✓	✓	1.0	65.9	58.9	**61.8**	37.5	1568.2	63.0

Table 2. Impact of training configurations with SigLIP + QWen2 on the reduced data mix. Applying our loss during pretraining improves performance over instruction tuning alone, while applying it at both stages yields the best results. The optimal configuration (α=0.1, applied at both stages) achieves superior performances across most benchmarks, demonstrating the effectiveness of our approach.

Encoder	LLM	Our Loss	Data	SEED	Real-World QA	CV-Bench^{2D}	MMVP	MME^P	GQA
SigLIP	QWen2	✗	Reduced	64.4	58.2	60.4	37.4	1535.8	62.6
SigLIP	QWen2	✓	Reduced	66.2	59.1	61.6	38.6	1578.0	63.5
CLIP	Vicuna	✗	Reduced	56.5	54.8	60.1	32.1	1464.5	61.2
CLIP	Vicuna	✓	Reduced	59.3	55.7	60.9	33.5	1529.0	62.4
SigLIP	QWen2	✗	Full	75.1	65.9	72.8	41.2	1601.6	65.2
SigLIP	QWen2	✓	Full	76.2	66.5	73.9	43.0	1629.7	65.7

Table 3. Evaluation of generalization across architectures and scalability to a larger dataset. Our method brings consistent improvements across different designs of MLLMs, and scales well to larger training data. When applying our loss to the full dataset yields superior performance across all benchmarks, confirming its effectiveness in large-scale MLLM training.

Layer index	SEED	Real-WorldQA	CV-Bench^{2D}	MMVP	MME^P	GQA
-	64.4	58.2	60.4	37.4	1535.8	62.6
7	64.7	58.4	60.7	37.2	1537.2	62.6
14	65.1	58.9	61.0	38.1	1565.7	63.0
21	65.9	58.2	60.5	37.6	**1582.3**	**64.1**
28	**66.2**	**59.1**	**61.6**	**38.6**	1578.0	63.5

Table 4. Impact of applying our loss function at different layer indices. Applying our novel objective in all chosen layers consistently boosts the performance, while applying it at a later layer yields better performance.

4.2. Architecture Analysis

To assess the integration of our proposed self-distillation loss into MLLMs, we conduct experiments using the reduced dataset, evaluating its effectiveness on perception-oriented benchmarks.

Recipe tuning. We first examine the impact of different training configurations, as shown in Table 2, considering both the stage at which the loss is introduced and the balancing term α (from Equation 8). Compared to the baseline, which does not apply our loss in either stage, all configurations incorporating our objective, except for the variant with a too-large balancing term, improve performance across all benchmarks. This reinforces both the effectiveness and robustness of our approach.

Next, we observe that applying the loss during pretraining yields better results than applying it solely during instruction tuning, likely because the alignment between visual and language spaces is primarily established in the pretraining phase. Our objective further reinforces this alignment, leading to more informative visual tokens. However, the best overall performance is achieved when the loss is applied at both stages. Additionally, a smaller balancing term produces better results, with $\alpha = 0.1$ achieving the best overall performance.

Choice of layers for applying our loss. We initially apply the reverse multimodal projector at the final self-attention layer of the LLM. To assess the impact of applying our loss function at different layer indices, we evaluate several variants, with results presented in Table 4. Our findings indicate that the objective consistently enhances performance across all settings, though its effectiveness is most pronounced when applied to later layers. Based on these results, we adopt the configuration of applying our loss at the last layer for all subsequent experiments.

With the optimal configuration identified, we assess the generalization and scalability of our self-distillation loss.

Generalization across architectures. To evaluate the architecture-agnostic nature of our method, we apply it to a widely used MLLM, LLaVA-1.5 [31], using the same reduced dataset (Table3, rows 3-4). As expected, this variant performs worse than our primary baseline, given that both its LLM and vision encoder are less expressive in multimodal reasoning. However, our method still yields con-

Model	Encoder	LM	Data Volume	SEED	Real-WorldQA	CV-Bench^{2D}	MMVP	MME^{P}	GQA
LLaVA-1.5 [32]	CLIP	VICUNA	1.2M	58.6	-	-	-	1510.7	62.0
MiniGemini-HD-8B [29]	CLIP	LLaMA3	2.7M	73.2	62.1	62.2	18.7	1606.0	64.5
LLaVA-NeXT-8B [33]	CLIP	LLaMA3	1.3M	72.7	60.1	62.2	38.7	1603.7	65.2
Cambrian-8B [46]	MoE	LLaMA3	10M	74.7	64.2	72.3	**51.3**	1547.1	64.6
LLaVA-One-Vision-7B [26]	SigLIP	QWen2	7.8M	75.4	66.3	-	-	1580.0	-
InternVL-8B [7]	Pretrain	Pretrain	6.1B	**76.2**	64.4	-	-	-	-
BLIP3 [54]	SigLIP	Pretrain	10B	72.2	60.5	-	-	1510.7	62.0
QWen2-VL-7B [50]	Pretrain	QWen2	Unknown	-	**70.1**	-	-	-	-
Grok-1.5 [52]	Unknown	Unknown	Unknown	-	68.7	-	-	-	-
Baseline	SigLIP	QWen2	8M	75.1	65.9	72.8	41.2	1601.6	65.2
Baseline^{+} (**Ours**)	SigLIP	QWen2	8M	**76.2**	66.5	**73.9**	43.0	**1629.7**	**65.7**

Table 5. Comparison with SOTA approaches on perception-oriented benchmarks. The top two results are marked in **bold** and underline. Our self-distillation-enhanced model ranks second among models trained on a similar data scale, trailing only LLaVA-One-Vision. On MME^{P}, it achieves SOTA performance among mid-sized MLLMs, validating the effectiveness of self-distillation in refining visual tokens.

Model	Encoder	LM	Data Volume	MMMU	TextVQA	ChartQA	LLaVA-bench	MMB	AI2D
LLaVA-1.5 [32]	CLIP	VICUNA	1.2M	-	58.2	-	65.4	64.3	-
MiniGemini-HD-8B [29]	CLIP	LLaMA3	2.7M	37.3	70.2	59.1	-	72.7	73.5
LLaVA-NeXT-8B [33]	CLIP	LLaMA3	1.3M	41.7	64.6	69.5	**81.6**	72.1	71.6
Cambrian-8B [46]	MoE	LLaMA3	10M	42.7	71.7	73.3	-	75.9	73.0
LLaVA-One-Vision-7B [26]	SigLIP	QWen2	7.8M	48.8	-	80.0	67.8	80.8	81.4
InternVL-8B [7]	Pretrain	Pretrain	6.1B	51.8	77.4	83.3	-	81.7	83.8
BLIP3 [54]	SigLIP	Pretrain	$\sim$ 10B	41.1	71.0	-	-	76.8	-
QWen2-VL-7B [50]	Pretrain	QWen2	Unkown	**54.1**	**84.3**	83.0	-	**83.0**	83.0
Grok-1.5 [52]	Unknown	Unknown	Unknown	53.6	78.1	76.1	-	-	88.3
LLaMA3.2-11B [9]	Pretrain	LLaMA3	6B	50.7	-	**83.4**	-	-	**91.1**
Baseline	SigLIP	QWen2	8M	43.1	72.3	74.0	67.5	78.1	79.1
Baseline^{+} (**Ours**)	SigLIP	QWen2	8M	43.6	71.8	72.1	69.2	78.3	79.6

Table 6. Core evaluation on general-purpose MLLM benchmarks. The top two results are marked in **bold** and underline. Our model also exhibits modest gains on several general-purpose benchmarks, though the improvements are less pronounced, suggesting that enhancing visual representations also positively impacts overall visual-language reasoning.

sistent improvements across all benchmarks, confirming its effectiveness across different architectures.

Scalability to larger dataset. Next, we scale our training from the reduced dataset to the full dataset, using our primary MLLM backbone. Results in Table 3 (rows 5-6) indicate that our self-distillation objective continues to provide improvements even at scale, demonstrating its ability to enhance MLLMs trained on large datasets. When trained on the full dataset, our method achieves superior performance across all evaluated benchmarks, reinforcing its robustness and scalability.

These findings suggest promising directions for further optimizing MLLM training strategies by preserving and enhancing visual representations at scale.

4.3. Comparison with the State-of-the-Art

We compare our final model against state-of-the-art mid-size models ($\leq$10B parameters), and present the results for perception-oriented and general-purpose benchmarks in Table 5 and 6, respectively. For the compared models, we use the performance metrics reported in their original papers. Our approach achieves competitive performance with other state-of-the-art models, demonstrating the effectiveness of enhancing visual token quality. The success of our method applied to this strong baseline model highlights the critical role of high-quality visual representations in MLLMs.

Perception-oriented benchmarks. Our self-distillation-enhanced model achieves top-tier performance, ranking second among state-of-the-art models trained on a similar data scale, trailing only LLaVA-One-Vision. Notably, our model consistently outperforms BLIP3, despite BLIP3 being trained on over 10B samples. On the MME^{P} benchmark, our model achieves state-of-the-art performance among mid-sized vision-language models, further validating the effectiveness of self-distillation in refining visual representations. These findings strongly reinforce our hypothesis that high-quality visual tokens are essential for MLLM performance, particularly in vision-centric tasks.

General MLLM benchmarks. Interestingly, our model

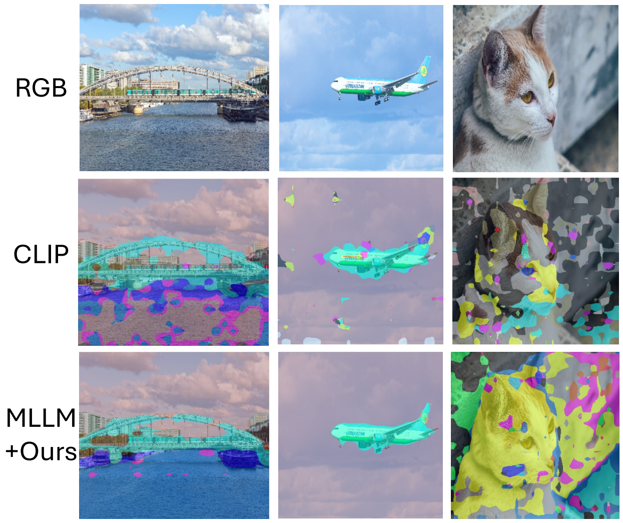

Figure 4. Visual comparisons for semantic segmentation between our MLLM model and its standalone CLIP encoder. The segmentations based on our MLLM visual tokens are smoother and more accurate, indicating that these visual tokens retain enough visual information thereby boosting the core vision-language reasoning capability of MLLMs.

also exhibits modest gains on several general-purpose benchmarks, though the improvements are less pronounced. This suggests that enhancing visual representations not only benefits perception tasks but also positively impacts overall visual-language reasoning. While the primary advantage of our approach lies in vision-centric applications, our findings indicate that improved visual tokens can contribute to broader MLLM capabilities.

4.4. Discussion

Visual Understanding Capability. In Section 2, we evaluated the visual tokens in MLLMs in comparison to their vision encoders on downstream tasks. Here, we extend this evaluation to all our models, with quantitative results shown in Figure 5. Across all models, the performance gap between the vision encoder and the MLLM output is significantly reduced. Notably, for the variant using the CLIP encoder, our model even surpasses the vision encoder in semantic segmentation, demonstrating that our approach effectively preserves critical visual information and enhances the vision-language reasoning capabilities of MLLMs. We further provide qualitative comparisons for this variant in Figure 4. Compared to its vision encoder, our final model produces visual tokens with clearer object boundaries and improved semantic consistency, further validating the effectiveness of our method.

Learning from an external visual expert model. Our primary evaluation follows a self-distillation approach, where final visual tokens are projected back into the image feature space using a cosine similarity loss. This idea can be extended by incorporating supervision from external visual expert models. In Table 7, we report results from variants that integrate CLIP as an auxiliary supervision source. Leveraging an external visual model further enhances MLLM performance, with mixture-of-experts (MoE)-based supervision yielding the best overall results compared to individual models. However, this approach introduces higher computational costs due to the additional supervision required during training. Alternative enhanced supervision methods, such as pixel-wise labels, also present promising directions, which we leave for future work.

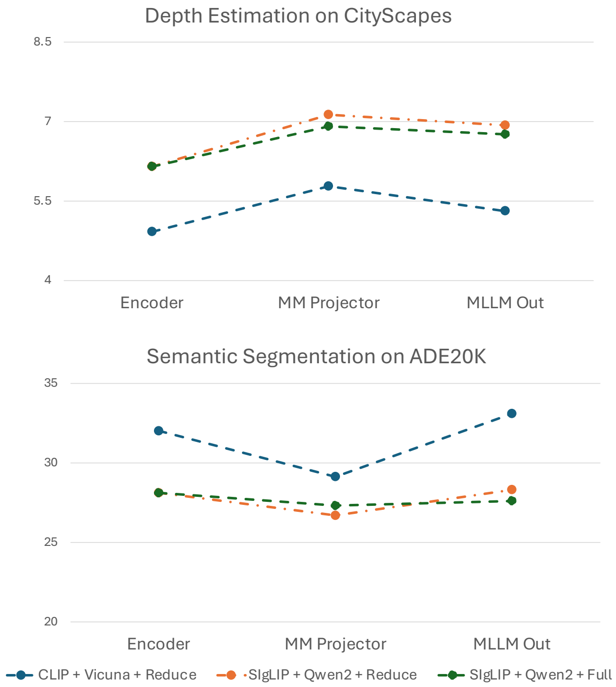

Figure 5. Quantitative evaluation of our method on downstream visual tasks. Applying our self-distillation objective enhances the intermediate visual tokens in MLLMs, leading to improved recognition capability.

Supervision Source	SEED	Real-WorldQA	CV-Bench^{2D}	MMVP	MME^{P}	GQA
CLIP	67.2	**59.7**	62.1	38.4	1564.8	64.1
SigLIP	66.2	59.1	61.6	**38.6**	1578.0	63.5
SigLIP + CLIP	**67.5**	59.6	**62.7**	**38.6**	**1588.1**	**65.2**

Table 7. Distilling from an external visual expert model. The results are obtained with SigLIP + QWen2 on the reduced data mix. Leveraging an external visual model further enhances MLLM performance, with mixture-of-experts (MoE)-based supervision yielding the best overall results compared to individual models.

Comparison with a heavier MM projector. Notice that our proposed reverse MM projector is not involved during the inference. However, to ensure that our improvements

Model	SEED	Real-WorldQA	CV-Bench^{2D}	MMVP	MME^{P}	GQA
Baseline	64.4	58.2	60.4	37.4	1535.8	62.6
4-layer Projector	64.8	57.7	60.5	37.0	1564.1	62.6
Baseline+ (**Ours**)	**66.2**	**59.1**	**61.6**	**38.6**	**1578.0**	**63.5**

Table 8. Comparison with a deeper four-layer MM projector. This model does not bring consistent improvements, confirming that the observed gains stem from our self-distillation approach.

are not merely due to additional training parameters, we introduce a baseline model with a deeper MM projector (increased from two to four layers) but without our self-distillation loss. As shown in Table 8, increasing the MM projector depth does not consistently improve performance over the standard version. This confirms that the observed gains stem from our self-distillation approach, rather than the inclusion of extra parameters during training.

Other discussions including the computational cost of our self-distillation loss, and the ablation study for our loss format are included in the supplementary material.

5. Related Work

Multi-modal Large Language Models (MLLMs) extend traditional large language models [9, 48, 55, 64] beyond solely processing natural language, integrating data from multiple modalities to enable a more comprehensive understanding and generation across diverse forms of information. Pioneering models like Flamingo and its successors [1, 3, 27] introduced visual adaptation layers (*e.g.* perceiver [18]) and cross-attention modules to fuse visual and language information. Building on this, models such as MM-GPT [13] and Otter [23], have leveraged well-constructed multimodal data to enhance conversational capabilities, expanding the utility of MLLMs as interactive chatbots with broader real-world applications. More recently, LLaVA [31] refined cross-attention architectures by projecting visual tokens into the language space and then processing them alongside language tokens within a pre-trained large language model. This method has shown promising improvements across diverse visual-language tasks. Continuing advancements in this domain, including efficient inference [4, 29, 66], video adaptation [11, 26, 58], and architecture-wise optimization [2, 6, 17, 32, 33, 45, 46, 50, 54, 57], bring MLLMs closer to real-world deployment and broaden their practical utility.

MLLMs for visual perception. While MLLMs excel in understanding natural images and generating language-based responses, recent research has explored extending their capabilities to visual grounding. Some approaches [5, 40, 51, 59, 62] focus on enabling MLLMs to engage in region-specific interactions, identifying and conversing about specific image regions (*e.g.*, bounding boxes or polygons). Although these models can handle region-focused data, their output remains text-based, limiting their effectiveness for visual grounding tasks that require more direct integration with visual data. In contrast, other methods have designed and trained MLLMs directly for visual perception tasks, such as referral segmentation [21, 42, 53] and images generation [12, 20, 39], by incorporating additional visual components to existing MLLM architectures. However, these methods typically rely on substantial, complex visual modules to enable such capabilities. By contrast, our approach aims to evaluate the core visual representations of MLLMs for visual understanding by applying lightweight task head, and further enhances their vision-language understanding capacity by refining these representations.

MLLM evaluation and benchmarks. The evaluation of MLLMs spans a wide range of tasks, including knowledge assessment [19, 56, 60], OCR [34, 36, 37], visual perception [10, 16, 24, 47, 52], *etc*. Beyond these fundamental evaluations, Beyond these core evaluations, Zhang et al. [63] examine the image classification capabilities of MLLMs, while Li et al. [25] investigates compositionality and biases within these models. POPE [28] highlights the issue of object hallucination in MLLMs and introduces a benchmark dataset to assess this problem. Cambrian [46] provides a comprehensive analysis of visual understanding in MLLMs, proposing a vision-centric benchmark for a thorough evaluation. In this work, we focus on enhancing the recognition capabilities of MLLMs, with particular emphasis on perception-related evaluations. Moreover, we also evaluate the representation of MLLM (visual tokens) on vision-specific datasets, including semantic segmentation on ADE20K [65] and depth estimation on Cityscapes [8]. Other potential datasets for evaluating MLLM visual tokens include image classification on ImageNet [43] and object detection on COCO [30].

6. Conclusion and Future Work

In this work, we introduce a novel self-distillation approach to enhance visual tokens in MLLMs, enabling more effective vision-language reasoning. Our method preserves crucial visual information through a reverse multimodal projection, consistently improving performance on recognition tasks while remaining adaptable across configurations. These findings highlight the importance of robust visual representations in MLLMs, opening pathways for future research in multimodal learning.

Future work. In this work, we supervise the visual tokens in MLLMs with features from vision encoders. A promising direction is to leverage additional visual data as supervision, *e.g.*, bounding boxes or dense pixel labels. We believe leveraging such data sources can potentially bring a larger boost to the recognition capacity of MLLMs and further improve the quality of language responses.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 8

[2] Anthropic. Claude 3.5, 2024. 8

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 8

[4] 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 a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 8

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 8

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 8

[7] 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 *CVPR*, 2024. 6

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2, 8

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6, 8

[10] 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. 4, 8

[11] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 8

[12] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 8

[13] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 8

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 4, 8

[17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 8

[18] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 3, 8

[19] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 4, 8

[20] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023. 8

[21] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 8

[22] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *ACL*, 2021. 2

[23] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 8

[24] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024. 4, 8

[25] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024. 8

[26] 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. 1, 6, 8

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 8

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 8

[29] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 6, 8

[30] 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*, 2014. 8

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2022. 1, 2, 3, 4, 5, 8

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2, 4, 6, 8

[33] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6, 8

[34] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 8

[35] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024. 4

[36] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 4, 8

[37] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 8

[38] 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. 2

[39] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 8

[40] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 8

[41] 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*, 2021. 2, 3, 4

[42] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 8

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 8

[44] 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*, 2019. 4

[45] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 8

[46] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 2, 4, 6, 8

[47] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 4, 8

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 8

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 8

[51] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 8

[52] XAI. Grok, 2024. 4, 6, 8

[53] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel-aligned language model. In *CVPR*, 2024. 8

[54] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 2, 6, 8

[55] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 4, 8

[56] Chengran Yang, Bowen Xu, Ferdian Thung, Yucen Shi, Ting Zhang, Zhou Yang, Xin Zhou, Jieke Shi, Junda He, DongGyun Han, et al. Answer summarization for techni-

cal queries: Benchmark and new approach. In *ASE*, 2022. 8

[57] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 8

[58] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 8

[59] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 8

[60] 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 *CVPR*, 2024. 4, 8

[61] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICV*, 2023. 3, 4

[62] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 8

[63] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*, 2024. 8

[64] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023. 2, 4, 8

[65] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 1, 2, 8

[66] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-ϕ: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024. 8