ViCToR: Improving Visual Comprehension via Token Reconstruction for Pretraining LMMs

Yin Xie[1*], Kaicheng Yang[1*], Peirou Liang[2*], Xiang An[1], Yongle Zhao[1], Yumeng Wang[1], Ziyong Feng[1], Roy Miles[3], Ismail Elezi[3], Jiankang Deng[4†]

[1]DeepGlint
[2]University of Science and Technology of China
[3]Huawei London Research Center
[4]Imperial College London
{yinxie,kaichengyang,xiangan}@deepglint.com, jiankangdeng@gmail.com

Abstract

Large Multimodal Models (LMMs) often face a modality representation gap during pretraining: while language embeddings remain stable, visual representations are highly sensitive to contextual noise (e.g., background clutter). To address this issue, we introduce a visual comprehension stage, which we call **ViCToR** (**Vi**sual **C**omprehension via **To**ken **R**econstruction), a novel pretraining framework for LMMs. ViCToR employs a learnable visual token pool and utilizes the Hungarian matching algorithm to select semantically relevant tokens from this pool for visual token replacement. Furthermore, by integrating a visual token reconstruction loss with dense semantic supervision, ViCToR can learn tokens which retain high visual detail, thereby enhancing the large language model's (LLM's) understanding of visual information. After pretraining on 3 million publicly accessible images and captions, **ViCToR** achieves state-of-the-art results, improving over LLaVA-NeXT-8B by 10.4%, 3.2%, and 7.2% on the MMStar, SEED^I, and RealWorldQA benchmarks, respectively.

Code — https://github.com/deepglint/Victor

Introduction

Large Language Models (LLMs) (Touvron et al. 2023; Achiam et al. 2023) have achieved remarkable success in text understanding and generation, driven by autoregressive Transformer architectures (Vaswani et al. 2023; Peng et al. 2025) and the scalability afforded by large-scale data and compute. However, these models remain fundamentally text-centric, limiting their applicability in multimodal contexts. To overcome this limitation, Large Multimodal Models (LMMs) have emerged, incorporating vision encoders such as CLIP (Radford et al. 2021) to transform images into token-like representations that LLMs can process. For instance, LLaVA (Liu et al. 2023b, 2024a; An et al. 2025) leverages GPT-4 (Achiam et al. 2023) to generate high-quality multimodal instruction-following datasets. Other approaches (Li et al. 2023b; Zhu et al. 2023; Wang et al. 2023) have introduced more advanced projection modules, such as Q-Formers and expert-based mechanisms, to improve vision-language alignment and task performance.

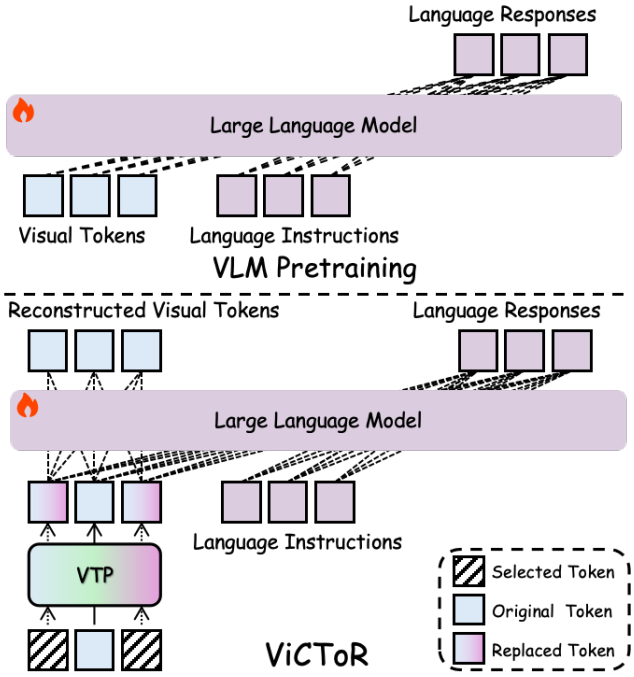

Figure 1: Traditional VLMs train language models to recognize visual tokens, while ViCToR instead replaces vision tokens with ones from a visual token pool, helping LLMs better understand and summarize images.

Notably, the VILA framework (Lin et al. 2024) pretrains with large-scale image-text interleaved data. However, these approaches suffer from relatively low efficiency. Recent insights suggest that using a relatively small amount of high-quality image captioning data, such as those from ShareGPT4v (Chen et al. 2024a), can lead to improved performance with a reduced computational cost. Another line of research (Jin et al. 2023) introduces novel visual tokenizers that convert non-linguistic images into discrete token se-

*Equal contribution
†Corresponding Author

quences, enabling LLMs to interpret them as foreign languages. Nevertheless, under such a unified framework, the intrinsic modality gap between vision and language remains unresolved, limiting the effectiveness of this approach.

To address these challenges, we introduce a visual comprehension stage and present **ViCToR**, a novel pretraining framework for LMMs. Specifically, we develop a learnable visual token pool (VTP) and adopt the Hungarian matching algorithm to select semantically relevant tokens from this pool for visual token replacement, as shown in Fig. 1. However, discretizing visual features into a limited set of tokens, while facilitating alignment with language, leads to severe loss of visual detail. To this end, we introduce a visual token reconstruction loss to maintain a faithful and effective visual representation. After pretraining on 3 million publicly accessible images and captions, **ViCToR** achieves state-of-the-art performance on multiple downstream benchmarks. In summary, our **contributions** are the following:

- We **design** a learnable visual token pool and employ the Hungarian algorithm to substitute selected original image tokens with the closest tokens from this pool.
- We **propose** a visual token reconstruction stage using a reconstruction loss and dense semantic supervision.
- We **demonstrate** that ViCToR achieves state-of-the-art performance on various benchmarks, surpassing LLaVA-NeXT-8B by up to 5.8%, showcasing strong capabilities in visual understanding and reasoning.

Related Work

Large Multimodal Model Pre-training. LLaVA uses a subset of the CC3M (Changpinyo et al. 2021) dataset for a more balanced coverage of concepts. Both the visual encoder and LLM are then frozen while the projection layer is trained to align the visual features and language tokens. However, relying solely on this approach will lead to a limited deep feature integration between the visual encoder and the LLM, mainly due to the restrictions imposed by the projection layer. To address this limitation, CogVLM (Wang et al. 2023) introduce a trainable visual expert module into the attention and feed-forward network layers of the language model. Despite this additional module, the LLM still remains limited by its frozen state and will continue to struggle interpreting the visual tokens. LaVIT (Jin et al. 2023) introduced a new visual tokenizer to convert images into a sequence of discrete tokens. However, directly inputting visual tokens into an LLM to enhance visual understanding with next-token prediction still presents significant difficulties. VILA (Lin et al. 2024) proposes an interleaved pretraining stage to augment the LLM to support visual input, but it relies on a 50M pretraining dataset, requiring considerable computational resources. Recent studies such as ShareGPT4V (Chen et al. 2024a) and LLaVA-OneVision (Li et al. 2024) demonstrate that high-quality image-caption pair data significantly improves the alignment between visual and textual modalities, thereby enabling more effective multimodal pretraining.

Visual Token Reconstruction. Masked Image Modeling (MIM) (He et al. 2021; Hondru et al. 2024) is now a common pre-training strategy for improving visual comprehension. Both 4M (Mizrahi et al. 2023) and MVP (Wei et al. 2022) propose to integrate this idea in the context of multimodal learning. In contrast, MILAN (Hou et al. 2022) proposes to reconstruct the image features infused with semantic content derived from caption supervision. Unmasked Teacher (Li et al. 2023d) selectively masks video tokens exhibiting low semantic content and aligns the remaining unmasked tokens through a linear projection to their counterparts from the teacher model. In a recent study, RILS (Yang et al. 2023) introduced a novel pre-training framework that employs masked visual reconstruction within a language semantic space. This framework facilitates the extraction of structured information by vision models through the accurate semantic prediction of masked tokens. Meanwhile, EVA (Fang et al. 2023) demonstrates that reconstructing the masked tokenized semantic vision features is an efficient strategy for vision-centric representation learning, removing the need for semantic feature quantization or any additional tokenization steps. There are also several important works, such as Ross (Wang et al. 2024) and Show-O (Xie et al. 2024), that leverage visual reconstruction tasks to train large multimodal models (LMMs). These approaches utilize large language models (LLMs) to reconstruct visual features, thereby enhancing both visual understanding and generation capabilities. Inspired by the above works, we propose a visual token reconstruction task for pretraining LMMs to bridge the modality gap between LLMs and visual tokens, enhancing the models' understanding of visual information.

Method

Preliminaries. Given an input image $I \in \mathbb{R}^{H \times W \times C}$, the vision encoder splits it into N patches p_i, each embedded as $z_i = E_v(p_i) \in \mathbb{R}^D$, forming a sequence z_i that encodes localized spatial information. In contrast, text tokens t_i are embedded using an embedding matrix $E \in \mathbb{R}^{|V| \times D}$, yielding $e_i = E(t_i)$; these discrete tokens are globally shared and derive relationships from co-occurrence in large corpora. This distinction leads to a modality representation gap.

While models like LLaVA (Liu et al. 2024c) attempt to align vision and language features via a learnable mapping, this process diverges from the LLM's original paradigm built for discrete language tokens. Consequently, the LLM only passively receives image features transformed to a language-like format, limiting its ability to fully capture the representational structure and inductive biases of visual data.

ViCToR. To overcome the limitations of existing pretraining paradigms, we propose a novel framework, ViCToR, for the pre-training of LMMs. In Sec. Visual Token Pool, we introduce the VTP for discretizing visual features into a limited set of tokens. In Sec. Visual Token Reconstruction, we then propose a visual token reconstruction task to recover the loss in visual detail from the token pool. Finally, in Sec. Dense Image Captioning Task, we show how to utilize the detailed captions to provide dense semantic supervision for vision token reconstruction. The complete training pipeline is outlined in Sec. Training Pipeline.

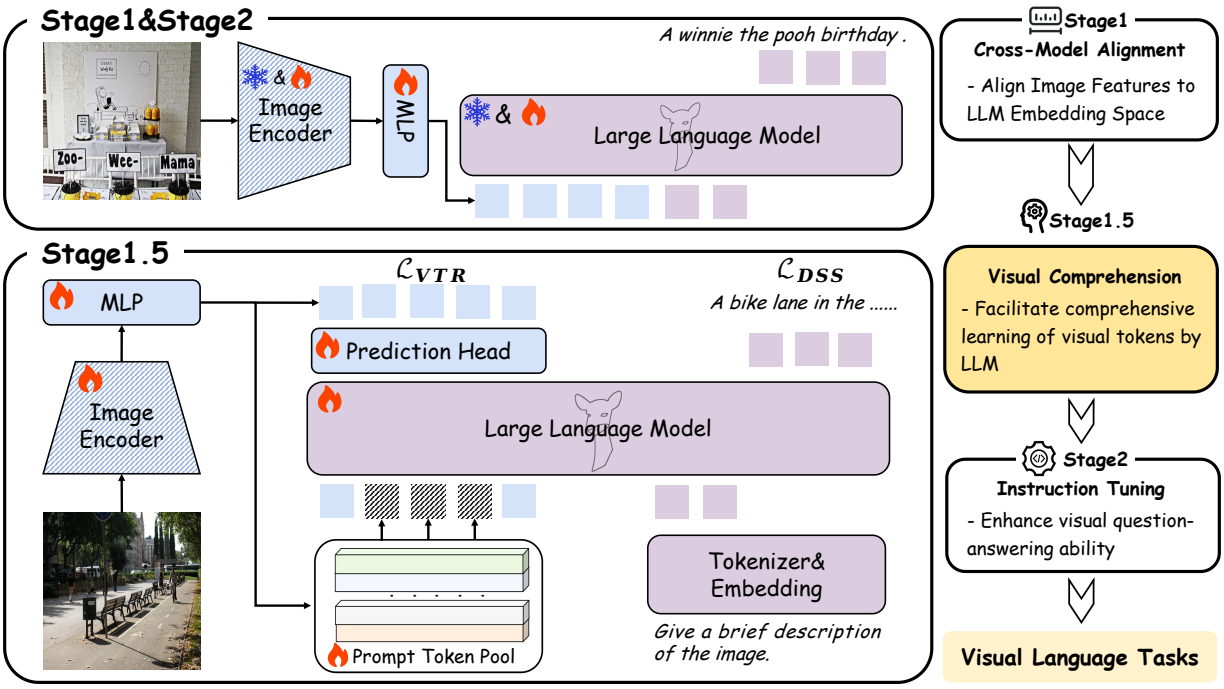

Figure 2: The training pipeline of our proposed ViCToR model. In contrast to LLaVA-1.5 (Liu et al. 2024b), we introduce an additional pre-training stage that involves visual token reconstruction and dense semantic supervision. This stage is essential for improving visual comprehension.

Visual Token Pool

To bridge the gap between continuous visual tokens and discrete language tokens, we introduce a VTP consisting of reusable and learnable visual tokens. Each selected token in a sequence of visual tokens is replaced with a learnable token selected from this pool. These tokens capture semantically meaningful visual patterns common across multiple samples, thereby enabling the LLM to actively learn the mapping between continuous tokens from the vision encoder and discrete tokens from the vision token pool within the language embedding space. We denote the VTP as $T_p \in \mathbb{R}^{N \times D}$, where N and D represent the number of learnable visual tokens and the feature dimension respectively. With a selection ratio of γ, we obtain the set of selected visual tokens $\widetilde{T}_v$. The selected tokens $\widetilde{T}_v$ are first padded with $\varnothing$ to maintain a consistent set size of N prior to assignment.

Token Assignment. We investigate two distinct methods for assigning tokens from the token pool. The first approach involves a simple nearest-neighbor (Cover and Hart 1967) lookup. However, this method often results in an over-reliance on a limited subset of tokens, leading to suboptimal utilization of the entire token space (refer to Sec. Ablation Study). Fortunately, this assignment issue is extensively addressed in combinatorics. The Hungarian algorithm (Kuhn 1955) provides a polynomial-time solution, with numerous efficient GPU adaptations available (Papadimitriou and Steiglitz 1998; Crouse 2023). Owing to its versatility, the algorithm has been widely adopted in various domains, including object detection (Carion et al. 2020).

Our objective is to determine a bipartite matching between $\widetilde{T}v$ and Tp that minimizes the total cost, defined here as the L2 distance between the replaced and original visual tokens. We address this by identifying a permutation of N elements, $\sigma \in \mathfrak{S}_N$, that minimizes this cost metric:

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \|\widetilde{T}_v^i - T_p^{\sigma(i)}\|_2 \quad (1)$$

The Hungarian algorithm addresses this assignment by iteratively subtracting the minimum values from each row and column. We observe that this optimization step is computationally inexpensive, constituting less than 5% of the total wall-clock time of the forward pass during training.

Visual Token Reconstruction

For this reconstruction task, we follow the same training paradigm of LLaVA, where an input image I is first encoded into a sequence of visual tokens. Specifically, visual features are extracted using a pretrained vision encoder (e.g., CLIP or SigLip2) E_v, and then projected into the language embedding space through a Multi-Layer Perceptron (MLP):

$$T_v = \{v_1, v_2, \ldots, v_n\} = \text{MLP}(E_v(I)) \in \mathbb{R}^{n \times d},$$

Method	Pub.	Res.	MMStar	RealWorldQA	MMBench$^{en}_{val}$	OCRBench	POPE	MMMU	AI2D	MME	SEED^{I}
LLaVA-1.5-13B	NeurIPS'23	336^2	34.3	55.3	67.8	337	**88.4**	37.0	61.1	1781	68.2
LLaVA-NeXT-8B	CVPR'24	672^2	43.9	58.4	–	531	87.1	43.1	72.8	<u>1908</u>	72.5
Cambrian-13B	NeurIPS'24	1024^2	47.1	<u>63.0</u>	<u>75.7</u>	<u>610</u>	86.8	41.6	73.6	1877	<u>74.4</u>
IDEFICS2-8B	NeurIPS'24	768^2	49.5	60.7	–	**626**	86.2	45.2	72.3	1848	71.9
Mantis-8B	TMLR'24	384^2	41.3	52.2	–	347	84.0	41.1	60.4	1675	68.5
Ross	ICLR'25	384^2	<u>53.9</u>	58.7	–	553	<u>88.1</u>	**49.0**	<u>79.4</u>	1854	73.6
ViCToR-7B		384^2	**54.3**	**65.6**	**79.0**	556	**88.4**	<u>48.9</u>	**79.5**	**2071**	**75.7**

Table 1: Comparison with other state-of-the-art vision-language models on various VLM benchmarks demonstrates that our method achieves leading performance across multiple domains. We highlight the best results in **bold** and the second-best results with an <u>underline</u>. All results of other methods reported in the tables are taken from their official papers and the Open VLM Leaderboard (Duan et al. 2024).

Method	Pub.	VE	LLM	Res.	Pretrain	Finetune	AI2D	MME	MMStar	RealWorldQA
LLaVA-NeXT-7B	CVPR'24	CLIP-L/14	Vicuna1.5-7B	672^2	558K	780k	67.0	1769	37.6	57.8
VILA1.5	CVPR'24	SigLIP-400M/14	LLaMA3-8B	384^2	50m	1m	58.8	1648	39.7	43.4
Sharegpt4V	ECCV'25	CLIP-L/14	Vicuna1.5-7B	336^2	558K+1.2m	742k	58.0	**1915**	35.7	54.9
ViCToR-7B		CLIP-L/14	Vicuna1.5-7B	336^2	558K+1.2m	780k	**70.9**	1873	**41.4**	**58.3**

Table 2: Comparison with other advanced pretraining methods for a fair evaluation. We mark the best performance **bold**. All results of other methods reported in the tables are taken from their official papers and the Open VLM Leaderboard (Duan et al. 2024).

where T_v denotes the sequence of n visual tokens, each of dimension d. For the visual reconstruction task, we randomly select a proportion γ of the visual tokens from T_v. Let $\mathcal{M} \subset \{1, 2, \ldots, n\}$ indicate the index set of the selected tokens, sampled uniformly at random such that $|\mathcal{M}| \approx \gamma n$. The selected token set is present as $\widetilde{T}_v = \{v_i \mid i \in \mathcal{M}\}$ and these tokens are replaced by visual tokens drawn from a VTP, resulting in a new visual input $\hat{T}_v$.

To supervise the reconstruction objective, for each index $i \in \mathcal{M}$, the model predicts a reconstructed visual token $\hat{v}_i$. These predictions form the reconstructed visual token set $\hat{T}_v = \{\hat{v}_i \mid i \in \mathcal{M}\}$. The reconstruction loss is then computed by measuring the distance between the predicted visual tokens $\hat{v}_i$ and the grounded visual tokens v_i for all selected positions. The loss of visual token reconstruction is defined as

$$\mathcal{L}_{\mathrm{VTR}} = \sum_{i \in \mathcal{M}} \|\hat{v}_i - v_i\|.$$

Dense Image Captioning Task

Captions which provide detailed visual semantic information can facilitate high-quality visual token reconstruction in LLMs. Thus, we introduce an additional task of detailed caption generation to help the model establish stronger associations between visual and language modalities.

Since visual token reconstruction encourages the LLM to actively model visual information by providing dense supervision, it is naturally complemented by a detailed caption generation task. This synergy helps the model form stronger associations between visual and linguistic modalities, leading to better generation capabilities. We minimize the loss of dense semantic supervision (DSS) by optimizing the negative log-likelihood of the target text tokens T_{rt}, conditioned on the visual context $\hat{T}_v$ and the previously generated tokens:

$$\mathcal{L}_{\mathrm{DSS}} = -\sum_i \log p\left(t_i \mid \hat{T}_v, t_1, \cdots, t_{i-1}\right). \quad (2)$$

Finally, the overall loss function is defined as:

$$\mathcal{L} = \alpha \mathcal{L}_{\mathrm{VTR}} + \mathcal{L}_{\mathrm{DSS}}, \quad (3)$$

where α is a loss weight to balance the influence of different losses.

Training Pipeline

Stage 1: Cross-modal Alignment. We begin by following LLaVA-1.5, pretraining the projection layer with 558K image-text pairs to align image features with the LLM embedding space. During this stage, the visual encoder and LLM are frozen, and only the projection layer is updated.

Stage 1.5: Visual Comprehension. To further improve cross-modal understanding, we introduce a visual reconstruction phase after alignment. This stage uses 3 million high-quality image-text pairs from the LLaVA-ReCap-CC3M dataset to jointly pretrain the visual encoder, projection layer, and LLM (Li et al. 2024).

Stage 2: Instruction Tuning. Finally, we perform instruction tuning on the whole model with 780K samples from

Prompt: <image>\nThe birds are swimming in a lake. Do you agree with my observation? Please state your judgment and provide reasons.

LLaVA-Next-8B: The image shows three birds that appear to be ducks or similar waterfowl species. They are in a body of water that looks like **a pond or a shallow lake**, given the presence of a paved area nearby and **the grassy bank**. The birds are wading through the water, which is typical behavior for ducks when they are foraging for food or moving from one area to another. **So, yes, I agree with your observation that the birds are swimming in a lake.**

ViCToR: **No, I do not agree with your observation.** The birds are not swimming in a lake; they are walking through **a puddle of water on the street**.

Figure 3: Qualitative Comparison of LLaVA-Next-8B and ViCToR-7B. Benefiting from our proposed cross-modal comprehension stage, the ViCToR model exhibits enhanced visual comprehension and reasoning capabilities. Moreover, it can generate enriched image descriptions.

LLaVA-NeXT, enhancing the model's ability to follow instructions and tackle various multimodal tasks (Liu et al. 2023a).

Experiments

Implementation Details. For our main experiments (Tab. 1), we use SigLIP2 So400m/14@384px (Tschannen et al. 2025) as the visual encoder and Qwen2.5-7B (Qwen et al. 2025) as the LLM. Additionally, we follow the LLaVA-1.5 setup by combining CLIP ViT-L/14@336px (Radford et al. 2021) with Vicuna (Chiang et al. 2023) 7B for comparison (Tab. 2).

In cross-modal alignment, the projection layer uses a learning rate of 1e-3. For visual comprehension, learning rates are set to 2e-5 for the LLM, projection layer, and VTP, and 2e-6 for the vision encoder. Key hyperparameters are a loss weight α of 10, random replace ratio γ of 75%, and VTP size of 2048. During instruction tuning, learning rates are 2e-5 for the LLM and projection layer, and 2e-6 for the vision encoder. AdamW (Loshchilov and Hutter 2019) is adopted with weight decay 0.2 and β_1/β_2 of 0.9/0.98. ViCToR is trained on 16× NVIDIA A800 GPUs for 35 hours.

Evaluation Benchmarks. We evaluate ViCToR across various benchmarks, including 1) OCR-Related Question Answering: OCRBench (Liu et al. 2024e); 2) hallucination: POPE (Li et al. 2023e); 3) Comprehensive Reasoning Benchmarks: MMBench (Liu et al. 2024d), RealWorldQA (xAI 2024), MMStar (Chen et al. 2024b), MME (Fu et al. 2024) and SeedBench^I (Li et al. 2023a); 4) Science Visual Question Answering: MMMU (Yue et al. 2024) and AI2D (Kembhavi et al. 2016).

Main Results

Tab. 1 presents a comprehensive comparison of our model with the LLaVA series, Cambrian (Tong et al. 2024), IDEFICS2 (Laurençon et al. 2024), Mantis (Jiang et al. 2024), and Ross across multiple benchmarks. Although other methods typically utilize higher input resolutions or larger training datasets, our model consistently achieves superior results on these benchmarks. Moreover, by using less training data, our approach not only significantly reduces

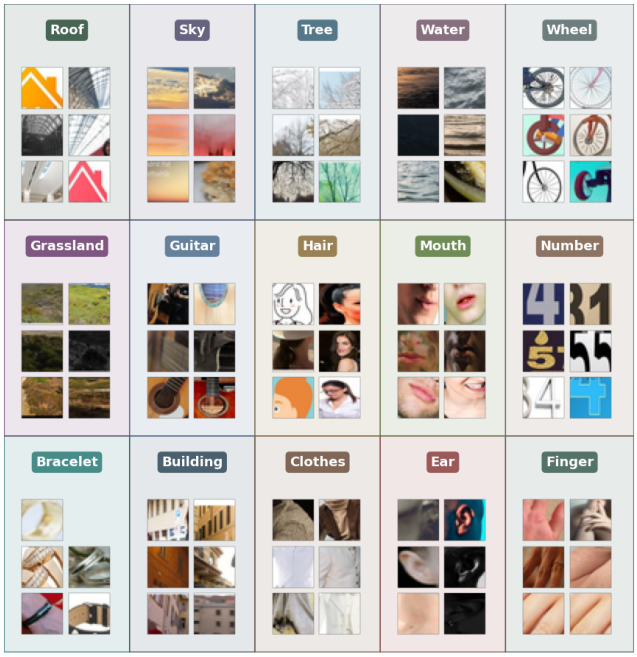

Figure 4: We select and visualize image regions consisting of more than four contiguous local patches that exhibit the shortest distance to the same item in the VTP.

training costs but also lowers inference costs due to the reduced input resolution. In addition, Fig. 3 showcases a number of examples highlighting the concrete differences in capabilities between our model and LLaVA-NeXT-8B.

Fair Comparative Experimental Designs. To provide an intuitive and rigorous evaluation of the effectiveness of our pre-training approach, we conduct comprehensive comparisons with other mainstream pre-training methods, including VILA and ShareGPT4V, under fair or even slightly disadvantageous conditions for our method. Specifically, we sample 1.2M pre-training examples, employ CLIP-L/14@336 as the vision encoder, and use Vicuna-1.5-7B as the language model. For baseline comparison, we select LLaVA-NeXT-7B with an image tiling strategy and ensure that its actual number of training tokens is at least comparable to ours. The

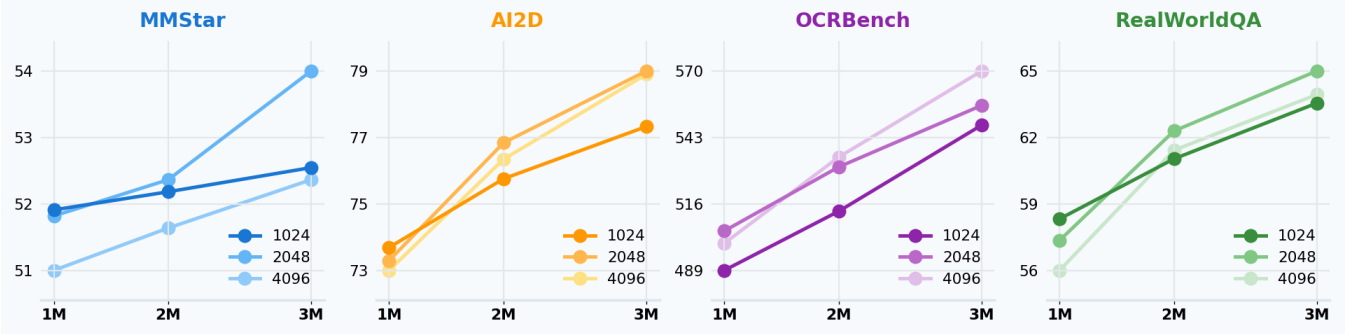

Figure 5: Performance comparison of different pre-training data scales and token pool sizes across various benchmark types.

final results, as shown in Tab. 2, demonstrate the comparative performance of our method.

Method	MMStar	AI2D	OCRBench	RealWorldQA
ViCToR$_{vq-gan}$	50.3	76.1	535	62.8
ViCToR$_{kmeans}$	52.8	76.3	548	61.3
ViCToR	**53.5**	**79.8**	**564**	**64.4**

Table 3: Comparison of different VTP initialization strategies on model performance across various benchmarks.

Insights and Analysis of the Visual Token Pool

We propose to use a VTP for bridging the modality gap between visual and textual tokens. We believe that this VTP enables the large language model (LLM) to easily aggregate region-level image features from all input images observed during training. To evaluate the classification capability of VTP-regional image features after pre-training, we group and visualize image patches that are closes to the same item in the VTP (see Fig. 4). The results clearly demonstrate that VTP can effectively cluster objects of the same category, even when they exhibit various visual forms, such as "wheel" and "roof".

What is the most effective initialization strategy for the VTP? We propose VTP to bridge the gap between linguistic and visual tokens, and initialize it randomly. VTP is then adaptively learned end-to-end during training under both vision and language supervision from the LLM. A natural question arises: does initializing VTP with visual features improve the LLM's visual understanding?

To test this, we pre-initialize VTP with visual features from 3M images using either VQ-GAN (Esser, Rombach, and Ommer 2021) reconstruction or K-means clustering, forming **ViCToR**$_{vq-gan}$ and **ViCToR**$_{kmeans}$. Performance is compared to random initialization (Tab. 3).

Results show that visual feature-based initialization does not offer improvements, and sometimes worsens performance. This supports our hypothesis that optimal VTP representations should be learned by the LLM, based on joint vision-language signals, rather than directly inheriting from the visual encoder.

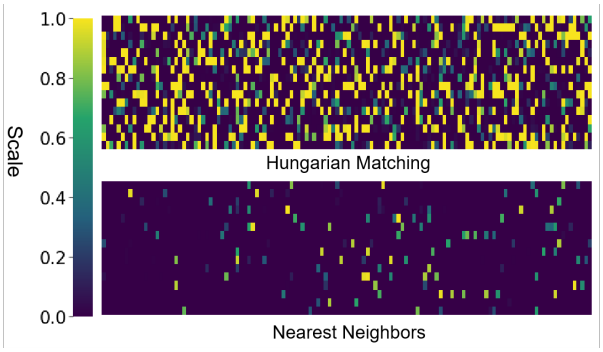

Figure 6: The token utilization rate in VTP with different matching algorithms with the ViCToR-7B model.

We further explore the effect of VTP size and pre-training data amount across four benchmarks. For three data scales, we test VTP sizes of 1024, 2048, and 4096. As shown in Fig. 5, increasing VTP size only improves performance when enough pre-training data is supplied. These findings highlight the necessity of scaling VTP size and pre-training data together for optimal results.

Ablation Study

Different reconstruction objectives and model architectures. In ViCToR, we use LLMs to reconstruct visual features for image understanding. We also test reconstructing pixels instead of features, keeping the MAE decoder and loss (He et al. 2022) and replacing the MLP adapter with QFormer (Li et al. 2023c). As Tab. 4a shows, pixel reconstruction, despite its low training loss, leads to worse downstream performance, suggesting that over-reliance on the decoder hinders efficient LLM comprehension of visual tokens. QFormer likewise offers no clear benefit.

Visual Token Replace Ratio. The replace ratio of visual tokens directly affects the difficulty of the visual token reconstruction task, and thus the effectiveness of pretraining. In Tab. 4b, we report the results of experiments with different mask ratios. Similar to the observations with MAE, we find that using a 75% replacement ratio yields optimal results in several downstream benchmarks. Lower ratios, e.g., 50%, make the pre-training task too easy, while higher ratios, e.g.,

Method	MME	OCRBench	POPE	SEED^I	RealWorldQA
ViCToR$_{pixel}$	1764	447	86.3	69.3	59.7
ViCToR$_{QFormer}$	1865	512	87.5	72.1	61.5
ViCToR	**2071**	**556**	**88.4**	**75.7**	**65.6**

(a) Effect of training objective.

Replace Ratio	MME	OCRBench	POPE	SEED^I	RealWorldQA
0.50	1908	532	**88.5**	74.2	60.1
0.75	**2071**	**556**	88.4	**75.7**	**65.6**
0.90	1978	528	88.0	74.8	63.4

(b) Effect of replace ratio.

Method	MME	OCRBench	POPE	SEED^I	RealWorldQA
Nearest Neighbors	1919	532	87.2	72.2	63.2
Hungarian Matching	**2071**	**556**	**88.4**	**75.7**	**65.6**

(c) Effect of matching algorithm.

S1	S1.5	MME	OCRBench	POPE	SEED^I	RealWorldQA
✗	✓	1884	514	86.3	72.8	63.1
✓	✓	**2071**	**556**	**88.4**	**75.7**	**65.6**

(d) Effect of removing Stage 1.

Table 4: Results of ablation studies on ViCToR-7B, evaluating the effects of different training objectives, replace ratios, matching algorithms, and the presence of Stage 1 across multiple benchmarks.

$\mathcal{L}_{VTR}$	$\mathcal{L}_{DSS}$	MMStar	RealWorldQA	MMBench$^{en}_{val}$	OCRBench	POPE	MMMU	AI2D	MME	SEED^I
✓	✗	50.3	63.2	75.4	516	**88.4**	48.3	78.1	1915	73.2
✗	✓	51.9	62.5	76.1	535	87.5	48.7	77.7	1847	73.9
✓	✓	**54.3**	**65.6**	**79.0**	**556**	**88.4**	**48.9**	**79.5**	**2071**	**75.7**

Table 5: Comparison of ablation results for $\mathcal{L}_{VTR}$ and $\mathcal{L}_{DSS}$ on multiple multimodal benchmarks.

90%, make it too difficult.

Nearest Neighbors v.s. Hungarian Matching. To improve the utilization rate of the tokens in the VTP, we use the Hungarian Matching. This is important to avoid an over-dependence on a small subset of the token pool, while also enabling a sufficient exploration of the VTP space. In Fig. 6, we present a comparative analysis of token utilization using both Nearest Neighbors and Hungarian Matching. Due to the Hungarian Matching requirement to select distinct visual tokens, we observe an improvement in the overall utilization of VTP, leading to a improvement across many evaluation benchmarks (Tab. 4c).

Stage 1 plays a critical role in our overall framework. Since the visual replace operation is performed after the visual features are processed by the adapter, Stage 1 is necessary to establish an initial alignment between the visual and linguistic spaces. Ablation studies (see Tab. 4d) demonstrate that omitting Stage 1 leads to a significant drop in performance and less efficient training, making it difficult for the model to achieve effective alignment.

Stage 1.5 Loss Ablation Study of ViCToR We conducted an ablation study on the two loss functions used in the Stage 1.5 phase of ViCoTR: the visual reconstruction loss($\mathcal{L}_{VTR}$) and the dense sequence supervision loss($\mathcal{L}_{DSS}$). Specifically, we systematically analyzed the effects of using each loss individually and in combination across nine different benchmarks, as shown in Tab. 5. The results demonstrate that $\mathcal{L}_{VTR}$ helps suppress visual hallucinations (e.g., on POPE) and excels in real-world scenario question answering tasks such as RealWorldQA, while $\mathcal{L}_{DSS}$ is more suitable for tasks requiring dense visual information understanding, such as OCR. Combining both losses achieves optimal performance on most benchmarks.

Conclusion

We present ViCToR, a novel pretraining methodology that significantly enhances the visual comprehension capabilities of Large Multimodal Models (LMMs). Our approach introduces a visual comprehension stage that effectively bridges the visual and textual domains, coupled with a dynamically learnable VTP leveraging the Hungarian algorithm for precise visual semantic processing. Extensive comparative experiments demonstrate that ViCToR outperforms state-of-the-art methods across multiple benchmark datasets, with particularly strong performance in visual understanding and cross-modal reasoning tasks. Through comprehensive ablation studies, we validate the efficacy and necessity of each component, with particular evidence supporting the critical contribution of the dynamic VTP to the model's overall performance. This work establishes a new paradigm for enhancing visual comprehension in multimodal foundation models and lays a solid foundation for the advancement of vision-language models.

Limitations

In this work, we have focused solely on image token reconstruction, which limits the scope to static images. However, for comprehensive video understanding, it is essential to consider both spatial and temporal token reconstruction. This would allow us to capture the dynamic changes that occur across frames and enhance the model's ability to process and interpret video sequences more effectively. Expanding our approach to include spatial-temporal token reconstruction is a necessary step for future improvements in video analysis.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

An, X.; Xie, Y.; Yang, K.; Zhang, W.; Zhao, X.; Cheng, Z.; Wang, Y.; Xu, S.; Chen, C.; Wu, C.; Tan, H.; Li, C.; Yang, J.; Yu, J.; Wang, X.; Qin, B.; Wang, Y.; Yan, Z.; Feng, Z.; Liu, Z.; Li, B.; and Deng, J. 2025. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. arXiv:2509.23661.

Carion, N.; Massa, F.; Synnaeve, G.; Nicolas Usunier, A. K.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024a. Sharegpt4v: Improving large multimodal models with better captions. In *European Conference on Computer Vision*, 370–387. Springer.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024b. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv:2403.20330*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27.

Crouse, D. 2023. On implementing 2D rectangular assignment algorithms. In *IEEE Transactions on Aerospace and Electronic Systems*.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11198–11201.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. arXiv:2012.09841.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377.

Hondru, V.; Croitoru, F. A.; Minaee, S.; Ionescu, R. T.; and Sebe, N. 2024. Masked Image Modeling: A Survey.

Hou, Z.; Sun, F.; Chen, Y.-K.; Xie, Y.; and Kung, S.-Y. 2022. Milan: Masked image pretraining on language assisted representation. *arXiv:2208.06049*.

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. MANTIS: Interleaved Multi-Image Instruction Tuning. arXiv:2405.01483.

Jin, Y.; Xu, K.; Chen, L.; Liao, C.; Tan, J.; Chen, B.; Lei, C.; Liu, A.; Song, C.; Lei, X.; et al. 2023. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv:2309.04669*.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A diagram is worth a dozen images. In *ECCV*.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? arXiv:2405.02246.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. arXiv:2307.16125.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023c. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, K.; Wang, Y.; Li, Y.; Wang, Y.; He, Y.; Wang, L.; and Qiao, Y. 2023d. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023e. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *CVPR*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024b. Improved baselines with visual instruction tuning. In *CVPR*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. arXiv:2304.08485.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024c. Visual instruction tuning. In *NeurIPS*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024d. Mmbench: Is your multi-modal model an all-around player? *ECCV*.

Liu, Y.; Li, Z.; Huang, M.; Yang, B.; Yu, W.; Li, C.; Yin, X.-C.; Liu, C.-L.; Jin, L.; and Bai, X. 2024e. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12).

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Mizrahi, D.; Bachmann, R.; Kar, O. F.; Yeo, T.; Gao, M.; Dehghan, A.; and Zamir, A. 2023. 4M: Massively Multimodal Masked Modeling. In *NeurIPS*.

Papadimitriou, C. H.; and Steiglitz, K. 1998. *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Peng, Z.; Huang, Y.; Xu, Z.; Tang, F.; Hu, M.; Yang, X.; and Shen, W. 2025. Star with Bilinear Mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25292–25302.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S. C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; Wang, Z.; Fergus, R.; LeCun, Y.; and Xie, S. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. arXiv:2406.16860.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

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.; and Zhai, X. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Wang, H.; Zheng, A.; Zhao, Y.; Wang, T.; Ge, Z.; Zhang, X.; and Zhang, Z. 2024. Reconstructive Visual Instruction Tuning. arXiv:2410.09575.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv:2311.03079*.

Wei, L.; Xie, L.; Zhou, W.; Li, H.; and Tian, Q. 2022. Mvp: Multimodality-guided visual pre-training. In *ECCV*.

xAI. 2024. Grok-1.5V. https://x.ai/news/grok-1.5v. Accessed: 2025-07-23.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. arXiv:2408.12528.

Yang, S.; Ge, Y.; Yi, K.; Li, D.; Shan, Y.; Qie, X.; and Wang, X. 2023. Rils: Masked visual reconstruction in language semantic space. In *CVPR*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv:2311.16502.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.