EM-KD: Distilling Efficient Multimodal Large Language Model with Unbalanced Vision Tokens

Ze Feng[1, 2*], **Sen Yang**[2], **Boqiang Duan**[2], **Wankou Yang**[1†], **Jingdong Wang**[2]
[1]Southeast University
[2]Baidu Inc
shannon.ze.feng@gmail.com, wkyang@seu.edu.cn

Abstract

Efficient Multimodal Large Language Models (MLLMs) compress vision tokens to reduce resource consumption, but the loss of visual information can degrade comprehension capabilities. Although some priors introduce Knowledge Distillation to enhance student models, they overlook the fundamental differences in fine-grained vision comprehension caused by unbalanced vision tokens between the efficient student and vanilla teacher. In this paper, we propose EM-KD, a novel paradigm that enhances the Efficient MLLMs with Knowledge Distillation. To overcome the challenge of unbalanced vision tokens, we first calculate the Manhattan distance between the vision logits of teacher and student, and then align them in the spatial dimension with the Hungarian matching algorithm. After alignment, EM-KD introduces two distillation strategies: 1) Vision-Language Affinity Distillation (VLAD) and 2) Vision Semantic Distillation (VSD). Specifically, VLAD calculates the affinity matrix between text tokens and aligned vision tokens, and minimizes the smooth L1 distance of the student and the teacher affinity matrices. Considering the semantic richness of vision logits in the final layer, VSD employs the reverse KL divergence to measure the discrete probability distributions of the aligned vision logits over the vocabulary space. Comprehensive evaluation on diverse benchmarks demonstrates that EM-KD trained model outperforms prior Efficient MLLMs on both accuracy and efficiency with a large margin, validating its effectiveness. Compared with previous distillation methods, which are equipped with our proposed vision token matching strategy for fair comparison, EM-KD also achieves better performance.

Introduction

Multimodal Large Language Models (MLLMs) (Bai et al. 2023b; Lu et al. 2024; Liu et al. 2024c; Li et al. 2023; Chen et al. 2024c) have emerged as a cornerstone of artificial intelligence, unifying visual and language understanding to tackle complex tasks such as visual question answering, image captioning, and embodied reasoning. While these models demonstrate remarkable versatility, their massive computational demands—stemming from intricate architectures and extensive multimodal pretraining—pose significant barriers to real-world deployment (Chen et al. 2024b; Xing et al. 2024; Zhang et al. 2025), particularly in resource-constrained environments. Recent efforts toward Efficient MLLMs (Yao et al. 2024; Li et al. 2025; Chen et al. 2024b; Feng et al. 2025b) aim to mitigate these challenges by re-

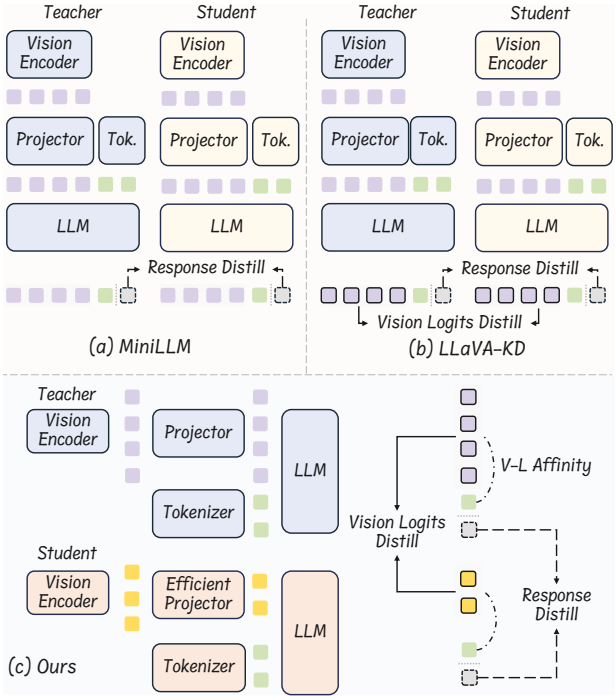


Figure 1: Comparison with existing MLLM distillation approaches. (a) MiniLLM (represent the LLM-style methods) focuses on response token distillation but neglects both the distinctive characteristics of visual features and vision-language correlations. (b) LLaVA-KD (represent the MLLM-style methods) aligns the teacher and student vision representations by modeling the correlations between vision tokens, but is limited to the condition of spatial alignment. (c) Our EM-KD can generalize to scenarios with unbalanced vision tokens between teachers and students, and introducing vision-language affinity distillation further enhances cross-modal alignment.

*This work was jointly done when ze feng was an intern in Baidu VIS.
†Corresponding author

moving the redundant vision tokens or compressing them in vision projectors, leading to higher computational efficiency. However, reducing vision tokens inevitably leads to information loss, resulting in degraded model performance and generalization, particularly in tasks requiring fine-grained image understanding (Feng et al. 2025b). Enhancing Efficient MLLMs comprehension without sacrificing inference efficiency remains a challenge that has not been fully explored.

As a training-phase-only technique, Knowledge Distillation can effectively improve student model capability and performance by incorporating guidance from teacher models or mimicking their behavioral patterns (Wang and Yoon 2021; Gou et al. 2021). Beyond its prevalent applications in CV and NLP, knowledge distillation has been successfully extended to MLLMs. LLaVA-KD (Cai et al. 2025) proposes a three-stage distillation paradigm which an MLLM is trained with instructions from a larger teacher. Align-KD (Feng et al. 2025a) employs an efficient teacher to distill an efficient student model, enhancing the student model's operational efficiency on edge devices. The difference between the teacher and student in previous methods lies solely in the size of models, and typically requires some MLPs to align their feature dimensions. In this work, we employ a well-pretrained vanilla teacher model to distill knowledge into an efficient student model, thereby enhancing the visual understanding and parsing capabilities of Efficient MLLMs. However, varying resolutions, vision encoders, and projectors lead to ***unbalanced vision tokens*** *between vanilla teacher and efficient student models* (shown in Figure 1). Previous approaches fail to generalize across such diverse configurations, as they inherently *rely on spatially aligned (i.e., one-to-one correspondences) vision token sequences.* Since the distillation object plays a critical role in the distillation strategy, here we conduct a preliminary analysis to identify an effective one. Inspired by (Neo et al. 2025), we employ the language model head (LM head) to decode vision tokens into vocabulary space (*i.e.* vision logits), revealing that image patches often map to semantically meaningful words and final-layer vision logits exhibit *rich semantics* (shown in Figure 2). Further, t-SNE visualizations of tokens (shown in Figure 3) demonstrate that vision and textual representations gradually intermix during LLM forward. Our analysis provides two directions: (1) it is feasible to treat vision logits as distillable semantic targets like text logits, and (2) distilling cross-modal relationships could further strengthen alignment between vision and language.

In this paper, we propose EM-KD, a novel paradigm that distills an Efficient MLLM with a well-pretrained teacher. To address the challenge of unbalanced vision tokens, we consider a set matching task and employ the Hungarian algorithm to establish optimal token-level correspondences between the teacher and student vision tokens. In order to achieve optimal token-wise distillation, we measure pairwise distances of vision logits to determine the best matches, followed by one-to-one knowledge transfer between aligned teacher-student pairs. Considering the semantic richness of vision logits in the final layer, EM-KD introduces Vision Semantic Distillation, which employs the reverse KL divergence to measure the distance between discrete probability distributions over the vocabulary space, rather than relying on hidden-level token similarity. Furthermore, we propose distilling the affinity matrix between text and aligned vision tokens (termed Vision-Language Affinity Distillation) to enhance cross-modal alignment.

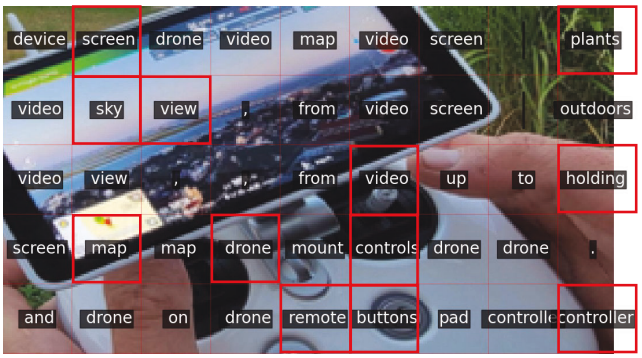
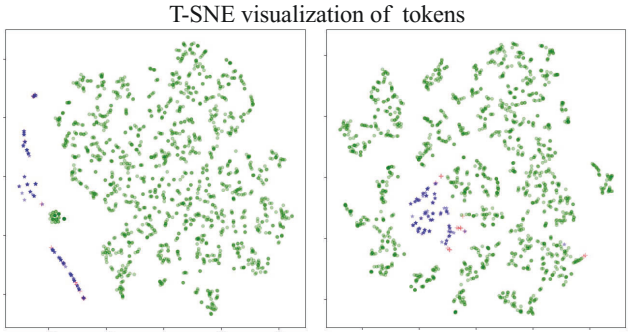


Figure 2: We decode each vision token into vocabulary space via LM head, and find vision logits exhibit rich semantic.



(a) Shallow layer (b) Deep layer

Figure 3: T-SNE visualization of tokens. Vision and textual representations gradually intermix during LLM propagation. Red: system tokens, blue: text tokens, green: vision tokens.

Comprehensive experiments on multiple vision understanding and parsing benchmarks indicate that EM-KD demonstrates performance improvements without altering the model architecture, validating the effectiveness. The EM-KD trained model outperforms previous Efficient MLLMs by a significant margin on both accuracy and efficiency, and when compared with previous MLLM distillation methods under broader evaluation settings, EM-KD consistently achieves superior performance. We summarize our contributions as follows:

- We propose EM-KD, a novel Knowledge Distillation framework designed to transfer visual comprehension and parsing capabilities from a vanilla teacher model to an Efficient MLLM.
- EM-KD overcomes the challenge of unbalanced vision tokens in the spatial dimension between the teacher and student with the Hungarian algorithm. For transferring more comprehensive knowledge, EM-KD intro-

duces Vision-Language Affinity Distillation and Vision Semantic Distillation.

- Compared with previous Efficient MLLMs, the EM-KD trained model exhibits superior performance on both accuracy and efficiency without altering the model architecture.

Related Work

Efficient Multimodal Large Language Models

By integrating vision encoders into the LLM (Touvron et al. 2023; Bai et al. 2023a; Liu et al. 2024a), MLLMs gain the capability for visual understanding and parsing (Alayrac et al. 2022; Liu et al. 2024c; Li et al. 2023). Since the computational complexity of Transformer is quadratic to the length of the token, these methods face significant challenges in inference efficiency.

Efficient MLLMs reduce computational cost by compressing or directly pruning redundant vision tokens. TokenPacker (Li et al. 2025) and Vision Remember (Feng et al. 2025b) mitigate vision forgetting caused by token compression through vision feature resampling mechanisms. Many methods focus on directly pruning redundant tokens without requiring post-training. FastV (Chen et al. 2024b) analyzes attention sparsity to identify redundant vision tokens and proposes early-stage pruning in shallow decoder layers. VisPruner (Zhang et al. 2025) strategically focuses on pruning vision encoders rather than LLM, preserving cross-modal alignment capabilities.

Knowledge Distillation in MLLMs

Many studies have introduced knowledge distillation techniques into the field of MLLMs. Based on the distillation targets, these approaches can be categorized into two types: LLM-style and MLLM-style. LLM-style methods (Lee et al. 2025; Ko et al. 2024, 2025; Wen et al. 2025; Gu et al. 2024; Shu et al. 2025) mainly focus on response token distillation. MiniLLM (Gu et al. 2024) adopts the Reverse Kullback-Leibler Divergence (RKLD) to avoid the student model overestimating the teacher's low-probability areas. Although LLM-style methods offer greater flexibility, they fail to fully unleash MLLMs' comprehension capabilities due to their neglect of both the distinctive characteristics of vision features and vision-language correlations. MLLM-style approaches (Xu et al. 2024; Cao et al. 2025; Cai et al. 2025; Feng et al. 2025a) primarily focus on distilling vision feature consistency and modeling vision-language correlations. LLaVA-KD (Cai et al. 2025) aligns teacher and student visual representations by modeling the correlations between vision tokens. However, LLaVA-KD can only perform distillation between teacher and student models that share identical vision encoders and projectors. Unlike these approaches, our proposed EM-KD targets a more general setting with unbalanced vision tokens.

Method

In this section, we first give a brief introduction to an Efficient MLLM, which serves as our baseline and student model. And then, we introduce the proposed EM-KD framework, including three key components: Vision Token Matching, Vision-Language Affinity Distillation, and Vision Semantic Distillation. The framework of EM-KD is shown in Figure 4.

Efficient Baseline

Representative MLLMs typically consist of three main components: (1) a vision encoder, (2) a vision projector, and (3) an LLM. The vision encoder, usually implemented as a Vision Transformer (Dosovitskiy et al. 2020) pretrained on large-scale datasets through vision-language contrastive learning approach, such as CLIP (Radford et al. 2021) and SigLip (Zhai et al. 2023), is responsible for extracting vision features from images. The currently widely used vision projector is an MLP, which aligns vision features to the language feature space. To reduce the number of vision tokens, we first compress vision features using Adaptive Average Pooling, followed by a two-layer MLP to form an efficient baseline, which also serves as our student model. One of the primary causes of unbalanced vision tokens is the different vision encoders in the teacher and student. In particular, the student model employs an efficient encoder with token compression, resulting less vision tokens, while teacher retain more vision tokens from powerful encoder to keep the fine-grained understanding capability. Finally, both vision tokens and text tokens are fed into a pre-trained LLM for processing and understanding. The MLLM then generates responses under the next-token prediction paradigm.

Vision Token Matching

As previously mentioned, the number of vision tokens differs between the teacher and student models in broader settings, and they are not one-to-one aligned in the spatial dimension. The misalignment fundamentally disrupts traditional MLLM-style distillation methods, which typically rely on strict token-level correspondence between teacher and student.

Given the vision tokens $T_v^t \in \mathbb{R}^{N_v^t \times D}$, $T_v^s \in \mathbb{R}^{N_v^s \times D}$ belonging to teacher and student respectively, we must find a permutation $\delta \in \Theta_{N_v^t}$ with the lowest cost:

$$\epsilon = \arg\min_{\delta} \sum \mathcal{C}_{match}(T_v^t, T_v^s), \tag{1}$$

where N_v^t, N_v^s indicate the vision tokens length in teacher and student, D is feature dimension, and $\mathcal{C}_{match}$ is the pairwise matching cost with permutation δ. Inspired by DETR (Carion et al. 2020), in EM-KD, we treat this as a set bipartite matching task and employ Hungarian algorithm. First, we decode the T_v^t, T_v^s into vocabulary space via the LM head θ^t, θ^s as the vision logits, and then we compute the *Manhattan distance* as the cost matrix $\mathcal{C}_{match} \in \mathbb{R}^{N_v^t \times N_v^s}$:

$$\mathcal{C}_{match} = Manhattan(\theta^t(T_v^t), \theta^s(T_v^s)). \tag{2}$$

Finally, the Hungarian algorithm running on the GPU is adopted to obtain the matching permutation δ. An alternative matching approach is to use 1D *Adaptive Average Pooling* to reduce the longer vision tokens. However, this method loses

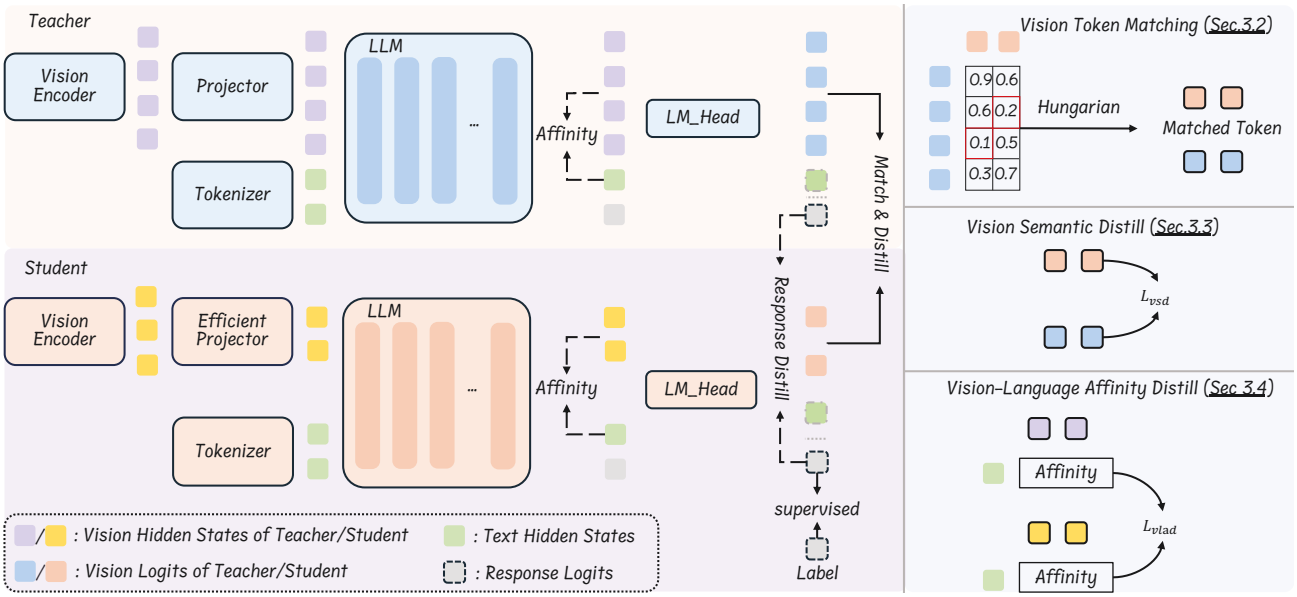


Figure 4: **Framework of the proposed EM-KD**. EM-KD consists of three key components: Vision Token Matching, Vision Semantic Distillation and Vision-Language Affinity Distillation. Vision Token Matching employs the Hungarian algorithm to resolve the unbalanced vision tokens problem between the teacher and student. Vision Semantic Distillation performs one-to-one knowledge transfer between matched tokens in logits space. Vision-Language Affinity Distillation further strengthens cross-modal alignment.

visual information and fails to leverage the teacher model's fine-grained understanding capability, leading to suboptimal accuracy (we will demonstrate in subsequent ablation experiments).

Vision Semantic Distillation

Considering the semantic richness (mentioned in the preliminary analysis in Sec.) of vision logits in the final layer, we propose using the *reverse KL Divergence* tailored for text logits to transfer the vision knowledge. Given the vision logits of the teacher and student models, and matching permutation δ, we calculate the loss function as follows:

$$\mathcal{L}_{\text{vsd}}(\pi_{\text{S}};\pi_{\text{T}}) = \mathbb{E}_{(x,y)\sim\pi_{\text{S}}}\left[\log\frac{\pi_{\text{S}}(y\mid x)}{\pi_{\text{T}}(y\mid x)}\right], \quad (3)$$

where $\pi_{\text{S}}(y \mid x)$ and $\pi_{\text{T}}(y \mid x)$ mean the probability distribution of the matched vision logits condition on input x for student and teacher. Distilling logits instead of hidden states eliminates the need for an additional projection layer to align the dimensions of the teacher and student models, as they share the same vocabulary. This is also a key benefit of Vision-Language Affinity Distillation.

Vision-Language Affinity Distillation

In traditional vision distillation methods, the primary focus is on modeling relation knowledge between vision features, and LLaVA-KD (Cai et al. 2025) also follows this principle. We argue that in MLLMs, the relationship between vision and language should be prioritized over vision feature relationships, as it directly measures vision-language alignment and evaluates the semantic affinity. Given the matched vision tokens $\hat{T}_v^t, \hat{T}_v^s \in \mathbb{R}^{N_v^s \times D}$ of the teacher and student models, we compute their *cosine distance* with the corresponding language tokens $T_l^t, T_l^s \in \mathbb{R}^{N_t \times D}$ as affinity matrices R_t, R_s:

$$R_t = Cos(\hat{T}_v^t, T_l^t),\ R_s = Cos(\hat{T}_v^s, T_l^s),\ \in \mathbb{R}^{N_v^s \times N_t}, \quad (4)$$

where D is the hidden dimension and N_t means the length of language tokens (in the SFT paradigm, they are equal in teacher and student). Then we compute the *Smooth L1* Loss between R_t and R_s:

$$\mathcal{L}_{\text{vlad}} = smooth_{L1}(R_t, R_s). \quad (5)$$

Although measuring affinity between logits (rather than hidden states) may seem like a more principled choice, the high dimensionality of logits would consume excessive GPU memory. We therefore adopt this memory-efficient alternative.

Overall Training Loss

Following the common practice, in addition to the vision semantic loss $\mathcal{L}_{\text{vsd}}$ and the vision-language affinity loss $\mathcal{L}_{\text{vlad}}$, we further incorporate supervised loss $\mathcal{L}_{\text{sup}}$ and distillation loss $\mathcal{L}_{\text{rld}}$ applied to the auto-regressive response logits as follows:

$$\begin{aligned}\mathcal{L}_{\text{sup}}(\pi_{\text{S}}) &= \mathbb{E}_{(y_k\mid y_{<k},x)\sim\pi_{\text{S}}}\left[\log\pi_{\text{S}}(y_k\mid y_{<k},x)\right],\\ \mathcal{L}_{\text{rld}}(\pi_{\text{S}};\pi_{\text{T}}) &= \mathbb{E}_{(x,y_k)\sim\pi_{\text{S}}}\left[\log\frac{\pi_{\text{S}}(y_k\mid y_{<k},x)}{\pi_{\text{T}}(y_k\mid y_{<k},x)}\right],\end{aligned} \quad (6)$$

The overall loss can be formulated as:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{sup}} + (1-\alpha)\mathcal{L}_{\text{rld}} + \beta\mathcal{L}_{\text{vsd}} + \gamma\mathcal{L}_{\text{vlad}}, \quad (7)$$

Listing 1: Pseudocode of EM-KD in a PyTorch-like style

```
# vhs_s, vhs_t: vision hidden states of
    student and teacher
# lhs_s, lhs_t: language hidden states of
    student and teacher
# rl_s, rl_t: response logits of student
    and teacher
# head_s, head_t: LM_head of student and
    teacher

L_disitll = as_tensor(0)
# decode the vision tokens as vision logits
vl_s, vl_t = head_s(vhs_s), head_t(vhs_t)
for data in batch:
  # no gradient when match vision logits
  with no_grad():
    # calculate Manhattan distance, same as
        Eq.(2)
    cost = cdist(vl_s, vl_t, p=1).detach()
    # Hungarian algorithm on GPU
    idx_s, idx_t = hungarian_gpu(cost).cpu
        ()

  # distill matched vision logits, same as
      Eq.(3)
  L_vsd = reverse KLD(vl_s[idx_s], vl_t[
      idx_t])

  # affinity between vision and text hidden
       states same as Eq.(4)
  affinity_s = cosine_similarity(vhs_s[
      idx_s], lhs_s)
  affinity_t = cosine_similarity(vhs_t[
      idx_t], lhs_t)

  # distill loss between affinity, same as
      Eq.(5)
  L_vlad = smooth_l1_loss(affinity_s,
      affinity_t)

  L_disitll += (beta * L_vsd + gamma *
      L_vlad)

# distill loss and sft loss between
    response logits, same as Eq.(6)
L_rld = reverse KLD(rl_s, rl_t)
L_sup = CrossEntropy(rl_s, labels)

loss = alpha * L_sup + (1 - alpha) * L_rld
    + L_disitll/batch_size
```

where α, β and γ are weights of each objective item to balance overall loss. Algorithm 1 provides the pseudo code of EM-KD in PyTorch-like style.

Experiments

Experimental Settings

Implementation Details. We choose the pretrained LLaVA-OneVision-SI (Li et al. 2024) in two different sizes (0.9B and 8B) as teacher models, and LLaVA-NeXT (Liu et al. 2024b) equipped with the efficient projector as student models. For the student model, we employ two scaled variants of Qwen2 (Yang et al. 2024) as the LLM paired with different-sized SigLip vision encoders (Zhai et al. 2023), constructing three baseline configurations with 0.6B and 8B parameters respectively. Following AnyRes strategy proposed in (Li et al. 2024), each image patch is processed by an efficient projector that compresses vision tokens to a maximum of 144 per patch.

We train the MLLM in two phases. *Phase-1: Language-Image Alignment.* In this phase, we use the image-caption pairs in the CC-558K dataset (Liu et al. 2024c) to train the efficient vision projector. *Phase-2: Visual Instruct Tuning.* In this phase, the whole student model is trained. The 779K mixture dataset (Liu et al. 2024b) is used to enhance the MLLM's capability of vision understanding and instruction following. We train all models in each phase for one epoch and the proposed EM-KD is only used in Phase-2. The loss weights α, β and γ are set to 0.5, 0.25 and 25 to balance the loss items. The experiments are conducted on 8 $\times$ NVIDIA H20 GPUs. Dur to the pages limitation, more model configuration and implementation details could be found in the *Supplementary Materials*.

Evaluation. We conduct extensive experiments on 11 benchmarks to validate the understanding and parsing capabilities of the proposed method. Specifically, we categorize all benchmarks into three distinct groups based on different focus areas: (1) General Question Answer benchmarks include GQA (Hudson and Manning 2019), MME-Perception (Fu et al. 2024) and RealWorldQA (xAI team 2024), (2) Comprehensive Knowledge Reasoning benchmarks include ScienceQA_Image (Lu et al. 2022), AI2D (Kembhavi et al. 2016), MMMU (Yue et al. 2024) and MM-Star (Chen et al. 2024a), (3) OCR&Chart Parsing benchmarks include ChartQA (Masry et al. 2022), DocVQA (Mathew, Karatzas, and Jawahar 2021), TextVQA (Singh et al. 2019) and OCRBench (Liu et al. 2023). To compare the accuracy of MLLMs, we take the average scores on the whole benchmark. For efficiency comparison, we evaluate the *Time to First Token (TTFT)* latency on the RealWorldQA.

Main Results

Comparison with other Efficient MLLMs. The ultimate goal of our work is to train an Efficient MLLM. To this end, we conduct comparisons with representative approaches for training-free vision token pruning (including FastV (Chen et al. 2024b), PyramidDrop (Xing et al. 2024), and VisPruner (Zhang et al. 2025)) and training-based vision token compression (including DeCo (Yao et al. 2024), TokenPacker (Li et al. 2025)). All methods are trained on the same datasets, and employ Qwen2-0.5B as the LLM and SigLip-base as the vision encoder.

For training-free vision token pruning approaches, we first train the LLaVA-NeXT (Liu et al. 2024b) model without pruning, serving as their baseline. And then employ the FastV (Chen et al. 2024b), PyramidDrop (Xing et al. 2024), and VisPruner (Zhang et al. 2025) to prune the vision tokens. For DeCo (Yao et al. 2024) and TokenPacker (Li et al. 2025), we directly replace the vision projector and train the

Methods	General			Knowledge				OCR&Chart				Average↑	TTFT/ms↓
	GQA	MME^P	RWQA	SQA^I	AI2D	$MMMU^V$	MMStar	ChartQA	DocVQA	TextVQA	OCRBench		
Vanilla Method without Vision Token Pruning or Compression													
LLaVA-NeXT (Liu et al. 2024b)	59.3	1203.7	48.2	53.1	52.3	30.2	34.3	57.6	61.9	47.3	38.6	**49.4**(△0.0)	**103.3**(△0.0)
Training-free Vision Token Pruning													
FastV (Chen et al. 2024b)	56.3	1200.3	47.9	53.0	51.0	31.0	33.2	51.5	57.4	43.5	31.6	**47.0**(-2.4)	**75.1**(+28.2)
PyramidDrop (Xing et al. 2024)	57.3	1210.4	48.1	53.1	51.3	29.8	34.4	52.3	58.4	45.0	33.9	**47.7**(-1.7)	**80.5**(+22.8)
VisPruner (Zhang et al. 2025)	57.1	1205.6	48.2	52.6	51.2	30.7	34.5	52.6	58.2	44.8	34.7	**47.7**(-1.7)	**61.7**(+41.6)
Training-base Vision Token Compression													
DeCo (Yao et al. 2024)	57.7	1157.6	50.0	53.0	51.4	30.7	34.1	53.8	57.4	44.3	34.4	**47.7**(-1.7)	**54.9**(+48.4)
TokenPacker (Li et al. 2025)	57.4	1201.5	47.7	51.8	51.3	29.6	34.3	52.9	56.7	41.4	33.5	**47.0**(-1.7)	**61.0**(+42.3)
EM-KD (Ours)	57.8	1226.5	51.5	53.1	51.5	31.0	36.1	59.1	64.1	49.2	39.7	**50.4**(+1.0)	**54.9**(+48.4)

Table 1: **Performance comparisons with various Efficient MLLMs.** When compared with other Efficient MLLMs, we all reduce the number of vision tokens to 144. We implement these methods under the same setting and test the TTFT on a single A100 GPU. The score of MME^P is divided by 20 when calculating the average accuracy. The results of our method are highlighted with blue. ↓ means that lower is better, ↑ means that higher is better. The best results are underlined.

Methods	General			Knowledge				OCR&Chart				Average↑
	GQA	MME^P	RWQA	SQA^I	AI2D	$MMMU^V$	MMStar	ChartQA	DocVQA	TextVQA	OCRBench	
LLaVA-OneVision-SI-0.5B distill 0.6B Efficient Baseline												
Efficient Baseline	58.7	1157.6	50.0	53.4	51.8	30.7	33.8	53.8	56.2	44.3	34.3	**47.7**
MiniLLM (Gu et al. 2024)	57.4	1247.3	51.5	53.9	51.4	29.0	35.0	57.3	62.4	47.4	36.8	**49.5**
LLaVA-KD* (Cai et al. 2025)	57.5	1203.8	52.0	52.6	52.6	29.9	34.5	57.3	62.7	48.2	39.0	**49.7**
EM-KD (Ours)	57.8	1226.5	51.5	53.0	51.5	31.0	36.1	59.1	64.1	49.2	39.7	**50.4**
LLaVA-OneVision-SI-7B distill 8B Efficient Baseline												
Efficient Baseline	61.4	1455.0	61.2	70.3	72.1	33.8	43.3	69.7	76.9	65.4	51.6	**61.7**
MiniLLM (Gu et al. 2024)	62.6	1478.4	61.8	51.0	71.7	71.7	38.3	45.1	70.4	77.0	64.2	**62.5**
LLaVA-KD* (Cai et al. 2025)	62.7	1465.3	61.8	72.1	71.5	38.5	45.3	69.5	75.1	63.9	51.3	**62.3**
EM-KD (Ours)	65.1	1514.9	62.2	77.4	72.4	38.9	46.0	69.6	74.3	64.1	51.7	**63.4**

Table 2: **Performance comparisons with various MLLM distillation approaches.** Efficient Baseline indicates the model was only trained with SFT supervision. * means we integrate the Vision Token Matching into LLaVA-KD. The score of MME^P is divided by 20 when calculating the average accuracy. We implement these methods under the same setting. The results of our method are highlighted with blue. ↑ means that higher is better.

model.

As shown in Table 1, the EM-KD trained model achieves 50.4 average score, outperforming all previous Efficient MLLMs by a significant margin, including FastV, PyramidDrop, TokenPacker and DeCo. An interesting finding is that our method still surpasses LLaVA-NeXT (49.4) with +1.0 accuracy improvement. Notably, LLaVA-NeXT is not an Efficient MLLM—it retains up to 576 tokens per image patch, resulting in higher latency, whereas EM-KD uses only up to 144 tokens per patch. This demonstrates that EM-KD produces models with fewer vision tokens (*i.e.*, more efficient inference) while achieving superior performance—fully validating the effectiveness of our methods.

Furthermore, we conduct efficiency comparisons between all methods. Compared with LLaVA-NeXT (Liu et al. 2024b) (103.3ms TTFT), which does not compress the vision tokens, all other methods achieve efficiency improvements due to the short embedding sequence. The acceleration effect of training-free vision token pruning methods is generally inferior to that of training-based vision token compression approaches. This discrepancy is because training-free methods rely on attention maps to select important tokens, which is fundamentally incompatible with mature acceleration operators such as Flash Attention and Scaled Dot-Product Attention. Among all methods, the model trained with EM-KD achieves the fastest speed (54.9ms), benefiting from the most compact network architecture.

Comparison with other knowledge distillation methods. To demonstrate the effectiveness of EM-KD, we make fair comparisons with two previous representative methods: MiniLLM (Gu et al. 2024) and LLaVA-KD (Cai et al. 2025). It is worth noting that LLaVA-KD cannot directly distill efficient student models due to the unbalanced vision tokens. We address this by integrating the same Vision Token Matching, which is also a key contribution in this work, in EM-KD to it.

As shown in Table 2, when using LLaVA-OneVision-SI-0.5B (actually 0.9B) to distill 0.6B efficient baseline, all dis-

	VLAD	VSD	RLD	General	Knowledge	OCR&Chart	**Average↑**
Baseline				55.5	42.4	47.2	**47.7**(△0.0)
EM-KD	✓			56.1	41.1	49.6	**48.4** (+0.7)
	✓	✓		56.6	42.1	51.5	**49.5** (+1.8)
	✓	✓	✓	56.9	42.9	53.0	**50.4** (+2.7)

(a) Ablation study of **key components**.

Matching Methods	General	Knowledge	OCR&Chart	**Average↑**
Average Pooling	55.8	41.5	50.2	**48.6** (-1.8)
Hungarian by Hidden States	56.6	41.8	51.5	**49.4** (-1.0)
Hungarian by Logits	56.9	42.9	53.0	**50.4**(△0.0)

(b) Experimental results with various **matching methods** in EM-KD.

Distill Objects of VSD	General	Knowledge	OCR&Chart	**Average↑**
Hidden States	56.4	42.1	51.6	**49.5** (-0.9)
Logits	56.9	42.9	53.0	**50.4**(△0.0)

(c) Experimental results with different **distillation objects of VSD** in EM-KD.

Table 3: Ablation studies. The default setting is marked in blue. ↑ means that higher is better.

tillation methods outperform the SFT-only baseline, validating our core hypothesis: MLLM distillation can not only perform well under varying model parameter scales, but also work across different vision token quantities. What's more, EM-KD achieves the highest average score, surpassing the baselines and other distillation methods. Specifically, EM-KD demonstrates consistent improvements across most tasks, such as DocVQA (64.1 *v.s.* 56.2 baseline), ChartQA (59.1 *v.s.* 53.8 baseline), and MMStar (36.1 *v.s.* 33.8 baseline). Compared to the LLM-style representative method MiniLLM (Gu et al. 2024), EM-KD surpass it by 0.9 (50.4 *v.s.* 49.5). This occurs because MiniLLM solely optimizes response tokens while disregarding the rich visual semantics information in vision tokens. Compared to the stronger competitor LLaVA-KD (Cai et al. 2025), because considering vision-language affinity, EM-KD yields a 0.7 point improvement in average score (50.4 *v.s.* 49.7). When we scale up the teacher and the student models, using LLaVA-OneVision-SI-7B to distill 8B efficient baselines, EM-KD exhibits consistent performance gains, outperforming two powerful counterparts.

Ablation Study

We conduct a number of ablation experiments to verify the effectiveness of each component in our EM-KD framework. Unless specified, we report the results on the 0.6B efficient student with the LLaVA-OneVision-SI-0.5B serving as teacher, and the experiment settings are kept as the same as the Sec .

Key Components. As shown in Table 3a. We evaluate the impact of three key components: VLAD, VSD, and RLD, on the performance across different task categories, including General, Knowledge and OCR&Chart, as well as the overall average score. Starting from the baseline, which achieves an average score of 47.7, we observe that introducing the VLAD alone leads to an improvement, raising the average score to 48.4 (+0.7). Adding the VSD on top of VLAD further improves the average score to 49.5 (+1.8 compared to baseline) When all three components (VLAD, VSD, and RLD) are combined, the model achieves the best performance with an average score of 50.4 (+2.7 compared to baseline). Significant improvements are observed both in General benchmarks (from 55.5 to 56.9) and OCR&Chart (from 47.2 to 53.0) benchmarks.

These results confirm that each component in our proposed EM-KD brings complementary benefits, and their joint application leads to the most substantial performance gains. The ablation study clearly demonstrates the effectiveness and necessity of all key components to achieve the best results.

Matching Methods. Table 3b presents the experimental results of different matching methods in the EM-KD, including Average Pooling, Hungarian matching by Vision Hidden States, and Hungarian matching by Vision Logits. Average Pooling achieves the lowest average score of 48.6. Hungarian matching by Vision Hidden States slightly improves the performance, reaching an average score of 49.4. When using Vision Logits to match the teacher and student, EM-KD achieves the best average score of 50.4. This experiment demonstrates that the Hungarian algorithm enables more precise matching, significantly enhancing the performance of distillation methods. Besides, matching based on Vision Logits is more beneficial because they possess explicit semantics, and distance metrics measured in vocabulary space give more accurate results than those in low-dimensional Hidden States.

Distillation Objects of VSD. Table 3c compares the effect of different distillation objects for VSD: (1) distilling Vision Hidden States and (2) distilling Vision Logits. When distilling Hidden States, the model achieves an average score of 49.5. Distilling Logits further improves the performance, yielding an average score of 50.4. This result suggests that using Logits as the distillation target is more effective than using Hidden States, consistent with the findings in the previous experiment. The vision logits encapsulate the distilled knowledge and lead to better overall generalization on all benchmarks, especially in the OCR&Chart dataset.

Conclusion

In this paper, we propose EM-KD, a novel paradigm that employs a well-pretrained teacher to distill Efficient MLLM. EM-KD addresses the visual token imbalance issue via the Hungarian algorithm, establishing strict token-level correspondence. Building upon the optimal assignment, we further introduce Vision-Semantic Distillation and Vision-Language Affinity Distillation to transfer more knowledge from the teacher to the student. We hope this work will draw the community's attention to Efficient MLLMs.

Acknowledgments

This work is supported by Shenzhen Science and Technology Program under project [JCYJ20230807114659029], Research Fund for Advanced Ocean Institute of Southeast University (Major Program MP202404), the National Natural Science Foundation of China under No. 62276061 and 62436002. This research work is also supported by the Big Data Computing Center of Southeast University.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Cai, Y.; Zhang, J.; He, H.; He, X.; Tong, A.; Gan, Z.; Wang, C.; and Bai, X. 2025. LLaVA-KD: A Framework of Distilling Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Cao, J.; Zhang, Y.; Huang, T.; Lu, M.; Zhang, Q.; An, R.; Ma, N.; and Zhang, S. 2025. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19846–19856.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

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

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024b. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, Q.; Li, W.; Lin, T.; and Chen, X. 2025a. Align-KD: Distilling Cross-Modal Alignment Knowledge for Mobile Vision-Language Large Model Enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4178–4188.

Feng, Z.; Liu, J.-J.; Yang, S.; Xiao, L.; Li, X.; Yang, W.; and Wang, J. 2025b. Vision Remember: Alleviating Visual Forgetting in Efficient MLLM with Vision Feature Resample. *arXiv preprint arXiv:2506.03928*.

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

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6): 1789–1819.

Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2024. MiniLLM: Knowledge Distillation of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 235–251. Springer.

Ko, J.; Chen, T.; Kim, S.; Ding, T.; Liang, L.; Zharkov, I.; and Yun, S.-Y. 2025. DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs. In *Forty-second International Conference on Machine Learning*.

Ko, J.; Kim, S.; Chen, T.; and Yun, S.-Y. 2024. DISTILLM: towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 24872–24895.

Lee, B.-K.; Hachiuma, R.; Ro, Y. M.; Wang, Y.-C. F.; and Wu, Y.-H. 2025. GenRecal: Generation after Recalibration from Large to Small Vision-Language Models. *arXiv preprint arXiv:2506.15681*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, W.; Yuan, Y.; Liu, J.; Tang, D.; Wang, S.; Qin, J.; Zhu, J.; and Zhang, L. 2025. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 1–19.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y.; Li, Z.; Yang, B.; Li, C.; Yin, X.; Liu, C.-l.; Jin, L.; and Bai, X. 2023. On the hidden mystery of ocr in large multimodal models. *arXiv e-prints*, arXiv–2305.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Sun, Y.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Mathew, M.; Karatzas, D.; and Jawahar, C. V. 2021. DocVQA: A Dataset for VQA on Document Images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Neo, C.; Ong, L.; Torr, P.; Geva, M.; Krueger, D.; and Barez, F. 2025. Towards Interpreting Visual Information Processing in Vision-Language Models. In *The Thirteenth International Conference on Learning Representations*.

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 *International conference on machine learning*, 8748–8763. PmLR.

Shu, F.; Liao, Y.; Zhang, L.; Zhuo, L.; Xu, C.; Zhang, G.; Shi, H.; Chan, L.; Yu, Z.; He, W.; et al. 2025. LLaVA-MoD: Making LLaVA Tiny via MoE-Knowledge Distillation. In *The Thirteenth International Conference on Learning Representations*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 preprint arXiv:2302.13971*.

Wang, L.; and Yoon, K.-J. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3048–3068.

Wen, Z.; Wang, S.; Zhou, Y.; Zhang, J.; Zhang, Q.; Gao, Y.; Chen, Z.; Wang, B.; Li, W.; He, C.; et al. 2025. Efficient Multi-modal Large Language Models via Progressive Consistency Distillation. *arXiv preprint arXiv:2510.00515*.

xAI team. 2024. Grok-1.5 Vision Preview.

Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; et al. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.

Xu, S.; Li, X.; Yuan, H.; Qi, L.; Tong, Y.; and Yang, M.-H. 2024. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yao, L.; Li, L.; Ren, S.; Wang, L.; Liu, Y.; Sun, X.; and Hou, L. 2024. DeCo: Decoupling Token Compression from Semantic Abstraction in Multimodal Large Language Models. *arXiv preprint arXiv:2405.20985*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang, Q.; Cheng, A.; Lu, M.; Zhang, R.; Zhuo, Z.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2025. Beyond Text-Visual Attention: Exploiting Visual Cues for Effective Token Pruning in VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.