Token-level Response-visual Attention Guidance for Multimodal LLMs Knowledge Distillation

Jaehyun Jang[1], Eunseop Yoon[1], Hee Suk Yoon[1], SooHwan Eom[1], Mark A. Hasegawa-Johnson[2], and Chang D. Yoo[1*]

[1] Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{jhjangjh,esyoon97,hskyoon,sean1105,cd_yoo}@kaist.ac.kr
[2] University of Illinois Urbana-Champaign, Champaign, USA
jhasegaw@illinois.edu

Abstract. While knowledge distillation (KD) is widely adopted for training lightweight models by leveraging supervision from larger teacher models, relying solely on output token distributions has proven insufficient for compressing Multimodal Large Language Models (MLLMs). Since output tokens are a byproduct of the model attending to visual inputs, prior works have explored explicitly distilling attention to provide a direct supervisory signal. While promising, the precise utility of which attention signals to distill remains under-explored. In this work, we challenge the conventional reliance on prompt-to-vision attention by revealing that downstream performance correlates strongly with response-to-vision attention similarity to the teacher, but negligibly with that of prompt-conditioned attention. Furthermore, we observe that attention distributions exhibit significant variance across individual tokens, indicating that a uniform distillation objective is suboptimal. To this end, we introduce **Token-level Response-visual Attention Guidance (TRAG)**, a distillation objective that 1) shifts the focus to response-to-vision signals and 2) employs token-specific objectives by adaptively weighting the Kullback-Leibler divergence based on attention entropy, effectively guiding the student to mirror the teacher's precise visual focus. Extensive experimental results on multiple benchmarks demonstrate that TRAG significantly outperforms prior distillation baselines. Our code is available at https://github.com/jhjangjh/TRAG.

Keywords: Multimodal Large Language Models · Knowledge Distillation · Attention Guidance

1 Introduction

Multimodal Large Language Models (MLLMs) [3, 22, 24, 26] have achieved remarkable performance, yet the continuous scaling of their architectures and datasets entails prohibitive computational costs. This deployment bottleneck has intensified research into MLLM compression [36, 44], where knowledge distillation (KD) [16] has emerged as a predominant paradigm, building upon its

* Corresponding Author

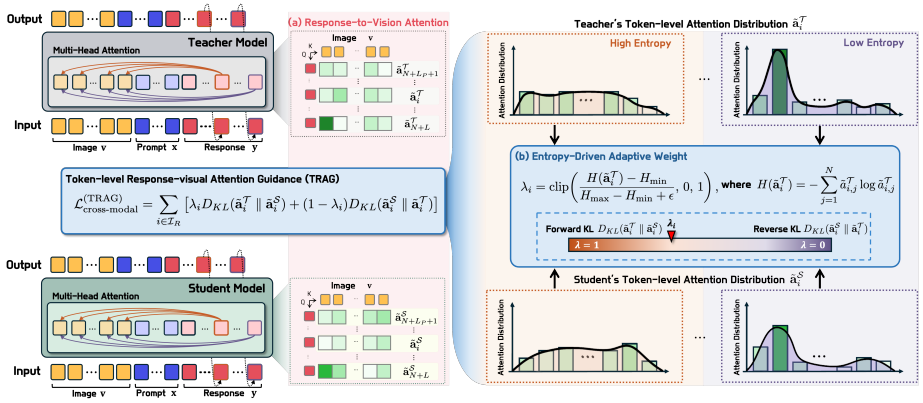

Fig. 1: Overview of Token-level Response-visual Attention Guidance (TRAG). (a) TRAG distills cross-modal grounding by isolating Response-to-Vision attention, anchoring the student's attention distribution ($\tilde{\mathbf{a}}_i^{\mathcal{S}}$) to the teacher's ($\tilde{\mathbf{a}}_i^{\mathcal{T}}$). (b) It dynamically modulates the distillation objective $\mathcal{L}_{\text{cross-modal}}^{(\text{TRAG})}$ for each response token via an entropy-driven adaptive weight λ_i, derived from the teacher's attention distribution entropy $H(\tilde{\mathbf{a}}_i^{\mathcal{T}})$. By shifting this weight toward the mean-seeking Forward KL for high-entropy distributions and the mode-seeking Reverse KL for low-entropy targets, TRAG ensures the student accurately mirrors the teacher's dynamic visual grounding during response generation.

established effectiveness in standard LLM frameworks [1, 12, 20]. However, standard distillation objectives designed for text-only LLMs rely solely on output distributions and thus do not explicitly transfer how visual evidence is grounded during response generation [19, 38, 40], even though output tokens are fundamentally a byproduct of the model attending to visual inputs.

To provide a more direct supervisory signal, prior works have explored explicitly distilling cross-modal attention [10, 19]. While this explicit guidance shows promise, identifying which attention signals to distill and at what granularity to apply supervision remains critically under-explored. Consequently, we find that the design choices made in existing frameworks often yield suboptimal supervision, failing to capture the dynamic evidence allocation that occurs during response decoding. In this work, we systematically characterize cross-modal attention distillation and its failure modes under existing design choices.

In detail, our contributions can be summarized as follows:

- **Response-to-Vision Attention as the Critical Signal:** We show that downstream performance correlates strongly with the alignment of Response-to-Vision attention between the student and the teacher model, whereas matching Prompt-to-Vision attention, which has been the primary focus of prior works, provides negligible utility. This identifies the generative phase, rather than the initial perception phase, as the critical signal for multimodal distillation.

- **Exposing the Limitations of Uniform Objectives:** We observe that cross-modal attention patterns exhibit significant variance across individual response tokens depending on their specific grounding roles, ranging from diffuse, scene-level context for function words to sharp, localized focus for visual entities. This high variance indicates that applying a uniform distillation objective across all tokens is suboptimal, necessitating an adaptive, token-wise supervision framework.
- Motivated by these findings, we propose **Token-level Response-visual Attention Guidance (TRAG)** (see Figure 1). TRAG shifts the distillation focus entirely to Response-to-Vision signals and introduces an adaptive, token-specific objective. By modulating the Kullback-Leibler (KL) divergence based on the teacher attention entropy, TRAG balances coverage and precision for each decoding step. Experiments on diverse multimodal benchmarks demonstrate that TRAG significantly outperforms competitive distillation baselines, achieving superior attention fidelity and robust performance gains.

2 Related Work

2.1 Knowledge Distillation in MLLMs

Prior KD approaches for MLLMs largely follow two trajectories. One focuses on improving modality-specific representation alignment and architectural adaptations. LLaVADI [38] analyzes key components for MLLM distillation, while LLaVA-KD [4] extends logit-based distillation to the vision modality and enforces structural consistency among visual tokens. Related compression efforts further improve compact MLLMs through visual-encoder mixtures [5] or MoE-based architectures [31]. While effective, these approaches do not explicitly supervise how visual evidence is selected during response generation, leaving visual grounding largely implicit.

Seeking more direct guidance on cross-modal processing, another trajectory explores attention-based distillation. Align-KD [10] transfers Prompt-to-Vision attention at a selected decoder layer, while CompoDistill [19] supervises Prompt-to-Vision attention maps. Concurrently, Align-TI [6] studies token-interaction distillation for multimodal alignment. However, these objectives mainly emphasize prompt-conditioned or globally defined interaction signals, overlooking token-wise dynamics during response generation. This gap motivates attention distillation that better aligns with response-time cross-modal processing, harmonizing with language modeling and logit distillation objectives.

2.2 Attention as a Distillation Signal

Attention has evolved into a key supervisory signal for knowledge distillation, serving as a medium to transfer both spatial focus and complex inductive biases. Early methodologies focused on establishing correspondence between attention maps or saliency patterns to guide the student's spatial perception

toward task-relevant visual cues [14, 42]. Subsequent developments introduced more structured forms of transfer, such as cross-modal structural constraints and sparse connectivity to highlight salient relationships [13, 30]. More recently, attention supervision has been integrated with representation transfer to better preserve the teacher's intermediate reasoning pathways and feature hierarchies [39]. These advancements establish attention as a robust mechanism for capturing the teacher's internal prioritization, providing the technical groundwork for specialized attention-based strategies in the multimodal domain.

3 Preliminaries

3.1 Cross-modal Attention Extraction

Standard MLLMs encode an input image into visual tokens $\mathbf{v} = \{v_1, \ldots, v_N\}$ via a vision encoder and an MLP-based projection module [24]. The model then processes the concatenated sequence $\mathbf{c} = [\mathbf{v}; \mathbf{x}; \mathbf{y}]$, where $\mathbf{x} = \{x_1, \ldots, x_{L_P}\}$ denotes the prompt token sequence and $\mathbf{y} = \{y_1, \ldots, y_{L_R}\}$ denotes the autoregressive response token sequence. Letting $L = L_P + L_R$, we index the elements c_i of $\mathbf{c}$ such that indices $i \in \{1, ..., N\}$, $i \in \{N+1, ..., N+L_P\}$, and $i \in \{N+L_P+1, ..., N+L\}$ map to the visual, prompt, and response tokens, respectively.

The generative process is driven by a Transformer decoder, whose self-attention at each layer $l \in \{1, \ldots, M\}$ produces an attention matrix $\mathbf{A}_l \in \mathbb{R}^{(N+L)\times(N+L)}$ with entries $a_{l,i,j}$ [34]. Under causal masking, each query index i can attend only to keys $j \leq i$. In particular, any textual query $i \in \{N+1, \ldots, N+L\}$ can attend to visual keys $j \in \{1, \ldots, N\}$ through cross-modal attention.

For any index set $\mathcal{I} \subseteq \{N+1, \ldots, N+L\}$ of textual queries, we define the cross-modal attention submatrix from $\mathcal{I}$ to visual keys as:

$$\mathbf{A}_l(\mathcal{I}) := \mathbf{A}_l[\mathcal{I}, 1{:}N] \in \mathbb{R}^{|\mathcal{I}|\times N}, \tag{1}$$

where $\mathbf{A}_l(\mathcal{I})[i, j] = a_{l,i,j}$ for $i \in \mathcal{I}$ and $j \in \{1, \ldots, N\}$. In this paper, we consider two instantiations of $\mathcal{I}$: Prompt-to-Vision attention uses $\mathcal{I}_P = \{N+1, \ldots, N+L_P\}$, whereas Response-to-Vision attention uses $\mathcal{I}_R = \{N+L_P+1, \ldots, N+L\}$.

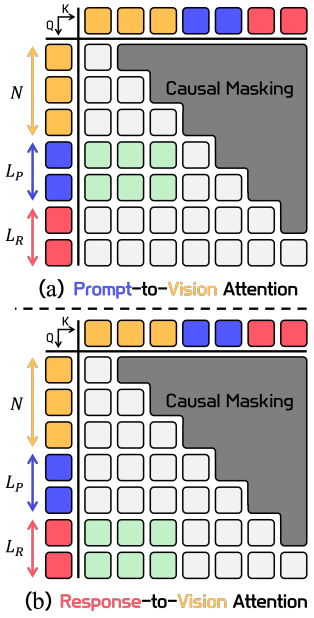

Fig. 2: Comparison of attention extraction according to query source.

- **Prompt-to-Vision Attention ($\mathbf{A}_{P\to V}$, Figure 2-(a)):** For each prompt token in $\mathbf{x}$ (i.e., c_i, where $i \in \mathcal{I}_P$), the distribution $\{a_{l,i,j}\}_{j=1}^{N}$ captures its attention over visual tokens during instruction encoding. This represents how

strongly the model integrates visual evidence when interpreting the prompt token.

- **Response-to-Vision Attention ($\mathbf{A}_{R\to V}$, Figure 2-(b)):** For each response token in $\mathbf{y}$ (i.e., c_i, where $i \in \mathcal{I}_R$), the distribution $\{a_{l,i,j}\}_{j=1}^{N}$ captures which visual tokens the model attends to when generating the next response token. This represents a token-level trace of how visual context influences the response distribution at position i.

3.2 Knowledge Distillation Objectives for MLLMs

In prior works, the distillation objective for transferring knowledge from a teacher model $\mathcal{T}$ to a student model $\mathcal{S}$ is defined as a composite loss consisting of language modeling, uni-modal distillation, and cross-modal attention distillation:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{uni-modal}} + \mathcal{L}_{\text{cross-modal}}. \tag{2}$$

Language Modeling Loss ($\mathcal{L}_{\text{LM}}$). The language modeling loss is the token-level cross entropy (negative log-likelihood) of the target response $\mathbf{y}$ conditioned on $\mathbf{v}$ and $\mathbf{x}$:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{L_R} \log p^{\mathcal{S}}(y_t \mid \mathbf{v}, \mathbf{x}, y_{<t}). \tag{3}$$

Uni-modal Knowledge Distillation Loss ($\mathcal{L}_{\text{uni-modal}}$). Uni-modal distillation aligns the teacher and student within each modality. This objective applies logit distillation for tokens in the response, alongside an auxiliary loss $\mathcal{L}_{\text{aux}}$:

$$\mathcal{L}_{\text{uni-modal}} = \sum_{t=1}^{L_R} D_{KL}\left(p^{\mathcal{T}}(\cdot \mid \mathbf{v}, \mathbf{x}, y_{<t}) \parallel p^{\mathcal{S}}(\cdot \mid \mathbf{v}, \mathbf{x}, y_{<t})\right) + \mathcal{L}_{\text{aux}}. \tag{4}$$

The first term performs response-level logit distillation over the target response sequence $\mathbf{y} = \{y_t\}_{t=1}^{L_R}$ [16] by minimizing the KL divergence $D_{KL}(\cdot \parallel \cdot)$ between the teacher and student next-token distributions at each generation step. The auxiliary loss $\mathcal{L}_{\text{aux}}$ introduces additional uni-modal supervision, typically defined on visual-side outputs or intra-visual relations. In LLaVA-KD [4], $\mathcal{L}_{\text{aux}}$ distills vision-token logits and matches affinity matrices among vision tokens. Similarly, CompoDistill [19] distills intermediate-layer visual self-attention by aligning vision-to-vision attention maps with a cosine-based loss.[3] Overall, $\mathcal{L}_{\text{aux}}$ strengthens modality-specific alignment by providing additional supervision beyond response-level logit distillation.

[3] We decompose the single attention distillation loss in CompoDistill [19] into uni-modal and cross-modal terms. The detailed derivation is provided in Appendix C.

Cross-modal Knowledge Distillation Loss ($\mathcal{L}_{\text{cross-modal}}$). Cross-modal distillation aligns the teacher and student by matching their cross-modal attention maps, where textual queries attend to visual tokens. In CompoDistill [19], this objective is instantiated with a group-based layer matching scheme to handle depth differences between the teacher and student:

$$\mathcal{L}_{\text{cross-modal}}^{(\text{CompoDistill})} = \sum_{l \in \mathcal{M}_S} \mathcal{D}_{\cos}\big(\mathbf{A}_{G_l}^{\mathcal{T}}(\mathcal{I}),\ \mathbf{A}_{l}^{\mathcal{S}}(\mathcal{I})\big)\,, \tag{5}$$

where $\mathcal{D}_{\cos}(\cdot,\cdot)$ denotes cosine distance, $\mathcal{M}_S$ denotes a subset of student layers selected for supervision, and $\mathcal{I}$ denotes the textual query indices used to extract $\mathbf{A}_l(\mathcal{I})$ (Eq. 1). For each student layer l, where G_l denotes the set of teacher layers assigned to l, the grouped teacher attention is

$$\mathbf{A}_{G_l}^{\mathcal{T}}(\mathcal{I}) := \frac{1}{|G_l|} \sum_{g \in G_l} \mathbf{A}_{g}^{\mathcal{T}}(\mathcal{I}). \tag{6}$$

In prior work, including Align-KD [10] and CompoDistill [19], $\mathcal{I}$ **is typically chosen as the prompt index range** $\{N+1, \ldots, N+L_P\}$ (i.e., they set $\mathcal{I}$ to $\mathcal{I}_P$), focusing supervision on Prompt-to-Vision attention.

4 Revisiting Cross-modal Knowledge Distillation

Building on Section 3, we revisit two common but under-discussed choices in attention-based cross-modal knowledge distillation. First, prior work often supervises attention on prompt tokens by default. Second, they typically devote substantial design effort to deciding which layers to distill from and how to align layers across teachers and students with different depths. *In this section, we ask whether prompt-based supervision is truly optimal, and whether layer selection is as critical as often assumed compared to the granularity of the attention signal.*

4.1 Is Aligning Prompt-to-Vision Attention to the Teacher Optimal for Knowledge Transfer?

If Prompt-to-Vision attention alignment to the teacher is an optimal distillation target, then Prompt-to-Vision attention similarity should consistently accompany a stronger student. To examine this, we compute teacher-student attention similarity and analyze its correlation with benchmark performance across multiple students. We measure teacher–student cross-modal attention similarity as cosine similarity under teacher forcing on an unseen dataset *VIGC-InstData* [35].

The results show that prompt-to-vision alignment does not reliably track model capability. As shown in Figures 3-(a) and (c), in Prompt-to-Vision attention ($\mathbf{A}_{P\to V}$), similarity to the teacher does not consistently correlate with downstream performance. Notably, the Supervised Fine-Tuning (SFT) baseline can achieve higher $\mathbf{A}_{P\to V}$ similarity than distilled models such as LLaVA-KD and CompoDistill, while still underperforming them on downstream benchmarks.

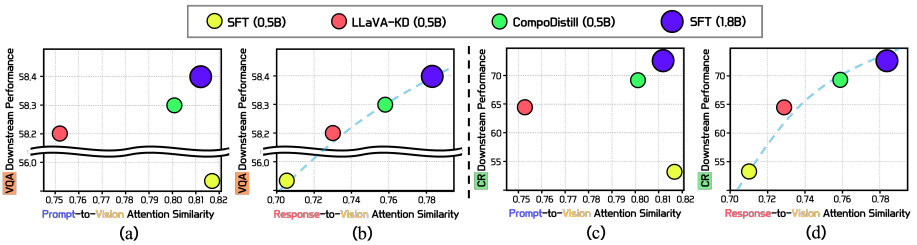

Fig. 3: Correlation analysis between attention similarity and downstream performance. (a) and (c) show that Prompt-to-Vision attention similarity exhibits negligible correlation with downstream performance on Visual Question Answering and Compositional Reasoning tasks. In contrast, (b) and (d) demonstrate a strong positive correlation between Response-to-Vision attention similarity and model performance, identifying the generative phase as the critical signal for multimodal distillation.

This indicates that aligning attention during prompt processing alone is not a sufficient proxy for the cross-modal behavior that distinguishes stronger models.

In contrast, Response-to-Vision alignment is substantially more predictive. Figures 3-(b) and (d) show a strong positive correlation between Response-to-Vision attention ($\mathbf{A}_{R\to V}$) similarity and performance across students. This suggests that the teacher's effective visual grounding is expressed more clearly during response generation than during prompt processing, motivating attention supervision that targets the generative phase.

4.2 Which Granularity Governs Knowledge Transfer: Layers or Tokens?

Section 4.1 suggests that the key transferable signal emerges during generation, where Response-to-Vision attention ($\mathbf{A}_{R\to V}$) tracks downstream performance. The next question is the granularity at which this signal should be distilled. Most attention distillation methods adopt layer-level granularity through layer selection or mapping. We ask whether this layer-level choice is necessary by analyzing $\mathbf{A}_{R\to V}$ across both decoder depth and individual response tokens.

Focusing on intermediate depths (relative depth in $[0.3, 0.6]$), where cross-modal interactions are most pronounced [7–9, 19, 29], Figure 4 shows that adjacent layers exhibit highly similar $\mathbf{A}_{R\to V}$ patterns, indicating substantial redundancy along depth. In contrast, $\mathbf{A}_{R\to V}$ varies sharply across response tokens. When generating function words (e.g., articles), the model exhibits weak visual grounding with uniform, blurry attention, whereas content tokens show more diverse patterns ranging from sharply localized evidence to diffuse, scene-level grounding. This suggests that the variation in cross-modal grounding is primarily token-dependent rather than layer-dependent, motivating token-wise supervision that adapts to each response token. Additional examples are provided in Appendix E.

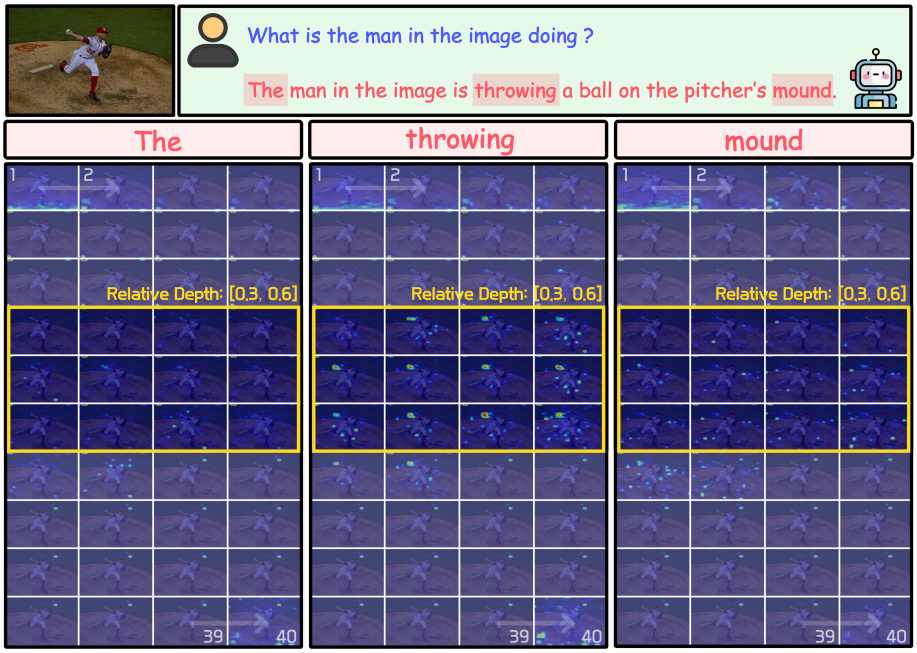

Fig. 4: Analysis of attention granularity across layers and tokens. Visualization of teacher attention maps reveals that while attention patterns exhibit high redundancy across adjacent layers, they show significant variance across individual response tokens. The yellow box marks intermediate layers (relative depth in $[0.3, 0.6]$ on a $[0, 1]$ normalized depth axis), where cross-modal interactions are most pronounced. The model exhibits weak, diffuse attention when generating the function word *'The'*, whereas it sharply concentrates on the pitcher's hand when generating *'throwing'* and shifts to the surrounding ground when generating *'mound'*.

5 Token-level Response-visual Attention Guidance

Motivated by our findings that Response-to-Vision attention alignment is a strong predictor of downstream task proficiency, and that cross-modal attention patterns vary substantially across response tokens (Section 4), we propose **Token-level Response-visual Attention Guidance (TRAG)**.

Building on the standard training objective (Eq. 2), we take the CompoDistill formulation (Eq. 5) as a starting point for the cross-modal distillation term, $\mathcal{L}_{\text{cross-modal}}$. TRAG reformulates this baseline into a response-token-wise objective that guides the student's Response-to-Vision attention using the teacher's attention as a reference for each target response token.

5.1 Shifting Supervision to the Response Phase

Motivated by the observation in Section 4.1, we shift cross-modal attention supervision from the prompt to the response, where visual grounding is expressed

during generation. Accordingly, we replace the prompt index set $\mathcal{I}_P$ with the response range $\mathcal{I}_R = \{N + L_P + 1, \ldots, N + L\}$:

$$\mathcal{L}_{\text{cross-modal}}^{(\text{response})} = \sum_{l \in \mathcal{M}_\mathcal{S}} \mathcal{D}_{\cos}\big(\mathbf{A}_{G_l}^{\mathcal{T}}(\mathcal{I}_R), \mathbf{A}_l^{\mathcal{S}}(\mathcal{I}_R)\big)\,. \tag{7}$$

This localizes attention supervision to response generation directly. Next, we refine the granularity of the distillation signal within this response range.

5.2 Adaptive Granularity via Entropy-Aware Weighting

Shifting from Global to Token-Level Supervision. Prior attention-based distillation methods typically define a single uniform objective over the supervised query set, comparing the teacher and student attention maps using a fixed discrepancy measure. However, Section 4.2 shows that token-level attention distributions are highly heterogeneous, suggesting that a single uniform constraint provides limited resolution to preserve per-token distributional intent. Moreover, Section 4.2 indicates substantial redundancy across intermediate layers, motivating aggregation across layers rather than sensitive layer selection.

Accordingly, we first aggregate cross-modal attention over intermediate layers to obtain a representative attention distribution for each response query, and then decompose supervision into token-level losses. Let $\mathcal{M}_\mathcal{S}$ and $\mathcal{M}_\mathcal{T}$ denote the sets of intermediate layers (relative depth in $[0.3, 0.6]$) used for aggregation in the student and teacher, respectively. For each response query index $i \in \mathcal{I}_R$, we denote the per-layer attention vectors over N visual tokens as $\mathbf{a}_{l,i}^{\mathcal{S}} := \{a_{l,i,j}^{\mathcal{S}}\}_{j=1}^{N}$ and $\mathbf{a}_{l,i}^{\mathcal{T}} := \{a_{l,i,j}^{\mathcal{T}}\}_{j=1}^{N}$, corresponding to the rows of $\mathbf{A}_l^{\mathcal{S}}(\mathcal{I}_R)$ and $\mathbf{A}_l^{\mathcal{T}}(\mathcal{I}_R)$ restricted to visual keys (Eq. 1). We then obtain layer-averaged attention vectors by averaging over intermediate layers $\mathcal{M}_\mathcal{S}$ and $\mathcal{M}_\mathcal{T}$:

$$\bar{\mathbf{a}}_i^{\mathcal{S}} := \frac{1}{|\mathcal{M}_\mathcal{S}|} \sum_{l \in \mathcal{M}_\mathcal{S}} \mathbf{a}_{l,i}^{\mathcal{S}}, \qquad \bar{\mathbf{a}}_i^{\mathcal{T}} := \frac{1}{|\mathcal{M}_\mathcal{T}|} \sum_{l \in \mathcal{M}_\mathcal{T}} \mathbf{a}_{l,i}^{\mathcal{T}}. \tag{8}$$

Since these vectors are obtained by restricting attention to visual keys, their total mass over $\{1, \ldots, N\}$ is not necessarily one. We therefore renormalize them on the vision span to form proper distributions:

$$\tilde{\mathbf{a}}_i^{\mathcal{S}} := \frac{\bar{\mathbf{a}}_i^{\mathcal{S}}}{\sum_{j=1}^{N} \bar{a}_{i,j}^{\mathcal{S}} + \epsilon}, \qquad \tilde{\mathbf{a}}_i^{\mathcal{T}} := \frac{\bar{\mathbf{a}}_i^{\mathcal{T}}}{\sum_{j=1}^{N} \bar{a}_{i,j}^{\mathcal{T}} + \epsilon}, \tag{9}$$

where $\epsilon > 0$ is a small constant for numerical stability. Finally, we define the token-level cross-modal objective as

$$\mathcal{L}_{\text{cross-modal}}^{(\text{token-level})} = \sum_{i \in \mathcal{I}_R} \mathcal{D}_{\text{guidance}}\big(\tilde{\mathbf{a}}_i^{\mathcal{T}},\ \tilde{\mathbf{a}}_i^{\mathcal{S}}\big)\,, \tag{10}$$

where $\mathcal{D}_{\text{guidance}}(\cdot, \cdot)$ is defined in the following subsection in detail.

Entropy-Driven Adaptive Weighting. Having decomposed the cross-modal objective into per-token terms (Eq. 10), we now derive a token-dependent coefficient that modulates the guidance behavior. For each response query index $i \in \mathcal{I}_R$, we compute the entropy of the teacher's normalized layer-aggregated attention distribution $\tilde{\mathbf{a}}_i^{\mathcal{T}}$ (Eq. 9):

$$H(\tilde{\mathbf{a}}_i^{\mathcal{T}}) = -\sum_{j=1}^{N} \tilde{a}_{i,j}^{\mathcal{T}} \log \tilde{a}_{i,j}^{\mathcal{T}}, \tag{11}$$

where $\tilde{a}_{i,j}^{\mathcal{T}}$ denotes the weight on the j-th visual token ($j \in \{1, \ldots, N\}$) in $\tilde{\mathbf{a}}_i^{\mathcal{T}}$.

We map $H(\tilde{\mathbf{a}}_i^{\mathcal{T}})$ to a bounded coefficient $\lambda_i \in [0, 1]$ via min–max normalization with clipping:

$$\lambda_i = \mathrm{clip}\left(\frac{H(\tilde{\mathbf{a}}_i^{\mathcal{T}}) - H_{\min}}{H_{\max} - H_{\min} + \epsilon}, 0, 1 \right), \tag{12}$$

where $H_{\min}$ and $H_{\max}$ are running lower and upper bounds updated online as exponential moving averages (momentum $\beta = 0.995$) of the mini-batch minimum and maximum of $H(\tilde{\mathbf{a}}_i^{\mathcal{T}})$, and $\epsilon > 0$ is a small constant for numerical stability. A larger λ_i corresponds to more diffuse attention, which we leverage to parameterize token-wise guidance in the following objective.

Distributional Guidance via Adaptive KL Regulation. To achieve specialized supervision, we define the guidance operator $\mathcal{D}_{\text{guidance}}$ as a weighted combination of forward and reverse Kullback-Leibler (KL) divergences. This formulation leverages the asymmetry of KL divergence. Forward KL, $D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{T}} \parallel \tilde{\mathbf{a}}_i^{\mathcal{S}})$, strongly penalizes the student when it assigns low probability to regions where the teacher has non-negligible mass, thereby promoting coverage. In contrast, reverse KL, $D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{S}} \parallel \tilde{\mathbf{a}}_i^{\mathcal{T}})$, penalizes probability mass placed outside the teacher's high-density regions more aggressively, which encourages concentration on dominant modes [2, 37].

Using the normalized coefficient λ_i defined in Eq. 12, we combine the two asymmetric KL terms into a token-wise guidance divergence per response token:

$$\mathcal{D}_{\text{guidance}}(\tilde{\mathbf{a}}_i^{\mathcal{T}}, \tilde{\mathbf{a}}_i^{\mathcal{S}}) = \lambda_i D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{T}} \parallel \tilde{\mathbf{a}}_i^{\mathcal{S}}) + (1 - \lambda_i) D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{S}} \parallel \tilde{\mathbf{a}}_i^{\mathcal{T}}). \tag{13}$$

Substituting Eq. 13 into Eq. 10 yields the final cross-modal objective (Figure 1):

$$\mathcal{L}_{\text{cross-modal}}^{(\text{TRAG})} = \sum_{i \in \mathcal{I}_R} \left[\lambda_i \, D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{T}} \parallel \tilde{\mathbf{a}}_i^{\mathcal{S}}) + (1 - \lambda_i) \, D_{KL}(\tilde{\mathbf{a}}_i^{\mathcal{S}} \parallel \tilde{\mathbf{a}}_i^{\mathcal{T}}) \right]. \tag{14}$$

This adaptive regulation overcomes the resolution limits of uniform objectives, enabling the student to capture visual evidence from localized to diffuse patterns.

5.3 Overall Training Framework

We adopt a three-stage training schedule (DPT–SFT–DFT), consistent with prior MLLM distillation works [4, 19]. DPT trains the projection module on

image–caption data with the language backbone frozen, SFT improves instruction following with the standard language modeling objective, and DFT reapplies teacher-guided distillation on the instruction data. The objective is stage-dependent: in DPT and DFT, we optimize

$$\mathcal{L}_{\text{TRAG}} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{uni-modal}} + \mathcal{L}^{(\text{TRAG})}_{\text{cross-modal}}, \tag{15}$$

whereas SFT minimizes only $\mathcal{L}_{\text{LM}}$. In Eq. 15, $\mathcal{L}_{\text{uni-modal}}$ follows the uni-modal distillation components of LLaVA-KD [4], and $\mathcal{L}^{(\text{TRAG})}_{\text{cross-modal}}$ is defined in Eq. 14. Detailed formulations are provided in Appendix G.

6 Experiments

6.1 Experimental Setup

Implementation Details. Both the teacher and student models use the same vision encoder, SigLIP-SO400M/14 at 384px resolution [43], and their LLM backbones are drawn from the Qwen1.5 and Qwen2.5 families. A two-layer MLP with GELU nonlinearity serves as the projection module to map visual features into the language embedding space [15].

We use *LLaVA-Pretrain-558K* for the DPT stage and *LLaVA-Instruct-665K* for the SFT and DFT stages [23]. Throughout all stages, the teacher model is frozen. For the student, we train only the projection module in the DPT stage, keeping the vision encoder and LLM frozen. In the SFT and DFT stages, we freeze the vision encoder and fully finetune the projection module and the LLM. All models are trained on 2 NVIDIA A100 GPUs with DeepSpeed ZeRO-2. Detailed hyperparameters are provided in the Appendix A.

Evaluation Benchmarks. We evaluate our models on two benchmark suites covering general Visual Question Answering (VQA) and compositional reasoning (CR). The general VQA suite consists of eight benchmarks, including GQA [18], ScienceQA [27], TextVQA [32], MME [11], MMBench in English and Chinese [25], POPE [21], and MMMU [41]. We further evaluate on three CR benchmarks: SugarCrepe [17], BiVLC [28], and Winoground [33]. Detailed evaluation metrics and setup are described in the Appendix B.

6.2 Main Results

General VQA benchmarks. Table 1 demonstrates that TRAG consistently improves performance over the PT-SFT baseline across all tested backbones and student scales. The performance gains are most significant in the $\leq$ 0.5B regime. For the Qwen2.5-0.5B student, TRAG increases Avg_6 from 62.4 to 65.3 and Avg_8 from 57.1 to 60.2 compared to the PT-SFT baseline. This result also surpasses LLaVA-KD on the same backbone by 1.2 points in Avg_6 and 1.5 points in Avg_8. Similar robust improvements are maintained for the Qwen1.5-0.5B backbone, where TRAG achieves the highest scores in both average metrics.

Table 1: General VQA benchmark comparison of TRAG. Performance on eight general VQA benchmarks across multiple backbones and parameter scales. Models are grouped by size; # Samples denotes the number of training samples. Avg_8 denotes the average over all 8 benchmarks, while Avg_6 averages the remaining 6 benchmarks excluding MMB^{CN} and MMMU. For the $\leq$ 2B and $\leq$ 0.5B categories, the best performing result is highlighted in **bold** and the second-best result is indicated with underline. Blue indicates the teacher model for each backbone, gray rows represent models trained with the standard PT-SFT pipeline, yellow denotes recent competitive distillation baselines, and green indicates our proposed method. The symbol † denotes results reproduced under our experimental setup for fair comparison.

Size	Method	LLM	# Samples	Visual Question Answering Benchmarks								Avg_6	Avg_8
				GQA	ScienceQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU		
≤ 4B	Bunny	Phi2-2.7B	2.6 M	62.5	70.9	56.7	74.4	68.6	37.2	-	38.2	-	-
	Imp-3B	Phi2-2.7B	1.5 M	63.5	72.8	59.8	-	72.9	46.7	-	-	-	-
	MobileVLM	MobileLLaMA-2.7B	1.2 M	59.0	61.0	47.5	64.4	59.6	-	84.9	-	62.7	-
	MoE-LLaVA	Phi2-2.7B	2.2 M	62.6	70.3	57.0	-	68.0	-	85.7	-	-	-
	MiniCPM-V	MiniCPM-2.4B	570 M	51.5	74.4	56.6	68.9	64.0	62.7	79.5	-	65.8	-
	TinyLLaVA	Phi2-2.7B	1.2 M	62.0	69.1	59.1	73.2	66.9	-	86.4	38.4	69.4	-
	TinyLLaVA†	Qwen1.5-4B	1.2 M	62.7	70.2	60.0	70.5	67.5	67.2	86.9	38.6	69.6	65.4
	TinyLLaVA†	Qwen2.5-3B	1.2 M	63.0	76.1	60.1	71.6	73.1	70.4	87.8	41.3	71.9	67.9
	LLaVADI	MobileLLaMA-2.7B	1.2 M	61.4	64.1	50.7	68.8	62.5	-	86.7	-	65.7	-
≤ 2B	Imp-2B	Qwen1.5-1.8B	1.5 M	61.9	66.1	54.5	65.2	63.8	61.3	86.7	-	66.3	-
	Bunny-2B	Qwen1.5-1.8B	2.6 M	59.6	64.6	53.2	65.0	59.1	58.5	85.8	-	64.5	-
	Mini-Gemini-2B	Gemma-2B	2.7 M	60.7	63.1	56.2	67.0	59.8	51.3	85.6	31.7	65.4	-
	MoE-LLaVA-2B	Qwen1.5-1.8B	2.2 M	61.5	63.1	48.0	64.6	59.7	57.3	87.0	-	63.9	-
	TinyLLaVA†	Qwen1.5-1.8B	1.2 M	60.9	65.2	47.7	61.3	57.8	56.2	83.4	34.8	62.7	58.4
	TinyLLaVA†	Qwen2.5-1.5B	1.2 M	61.7	70.9	58.4	70.4	70.6	64.0	86.0	**39.2**	69.7	65.1
	KD†	Qwen1.5-1.8B	1.2 M	60.9	64.6	53.3	66.3	65.8	63.1	86.5	32.7	66.2	61.6
	LLaVADI	MobileLLaMA-1.4B	1.2 M	55.4	56.0	45.3	58.9	55.0	-	84.7	-	59.2	-
	LLaVA-MoD	Qwen2-1.5B	5 M	58.8	69.2	59.9	69.2	68.9	64.4	**87.2**	-	68.8	-
	Align-KD	MobileLLaMA-1.4B	3.6 M	60.1	67.7	53.1	65.1	57.5	-	87.0	-	65.0	-
	CompoDistill†	Qwen1.5-1.8B	1.2 M	61.2	66.5	53.5	67.0	64.5	63.0	85.5	34.1	66.4	61.9
	LLaVA-KD	Qwen1.5-1.8B	1.2 M	62.3	64.7	53.4	69.1	64.0	63.7	86.3	33.6	66.6	62.1
	LLaVA-KD	Qwen2.5-1.5B	1.2 M	62.5	71.6	59.7	70.0	**71.0**	66.6	86.7	35.8	70.2	65.4
	TRAG	Qwen1.5-1.8B	1.2 M	61.7	66.9	55.6	67.0	66.6	63.3	86.8	35.2	67.4	62.9
	TRAG	Qwen2.5-1.5B	1.2 M	**62.6**	**73.2**	**60.0**	**71.2**	70.8	**69.3**	**87.2**	38.7	**70.8**	**66.6**
≤ 0.5B	TinyLLaVA†	Qwen1.5-0.5B	1.2 M	56.7	60.9	46.6	59.4	55.2	51.8	83.9	31.9	60.4	55.8
	TinyLLaVA†	Qwen2.5-0.5B	1.2 M	58.6	61.2	47.9	62.3	59.3	52.4	85.1	30.4	62.4	57.1
	KD†	Qwen1.5-0.5B	1.2 M	56.8	60.4	46.1	58.9	54.7	49.7	84.9	32.3	60.3	55.5
	LLaVA-MoD	Qwen2-0.5B	5 M	56.6	61.1	**57.1**	67.0	58.7	54.1	-	-	-	-
	CompoDistill†	Qwen1.5-0.5B	1.2 M	59.1	61.1	48.2	63.8	59.6	54.9	85.7	**34.0**	62.9	58.3
	LLaVA-KD	Qwen1.5-0.5B	1.2 M	59.6	60.6	49.9	64.5	60.1	55.5	85.9	30.2	63.4	58.2
	LLaVA-KD	Qwen2.5-0.5B	1.2 M	59.8	60.6	52.0	64.7	61.3	57.0	86.4	28.3	64.1	58.7
	TRAG	Qwen1.5-0.5B	1.2 M	59.5	62.1	48.9	65.2	60.9	54.6	86.5	30.9	63.8	58.5
	TRAG	Qwen2.5-0.5B	1.2 M	**60.2**	**62.7**	52.0	**68.1**	**61.7**	**57.7**	**87.6**	31.3	**65.3**	**60.2**

In the $\leq$ 2B regime, TRAG continues to exhibit superior performance. Using the Qwen2.5-1.5B backbone, TRAG achieves 70.8 in Avg_6 and 66.6 in Avg_8, outperforming both the TinyLLaVA baseline and LLaVA-KD. Notably, TRAG surpasses LLaVA-MoD (68.8 in Avg_6) despite using 76% fewer training samples (1.2M vs. 5M).

Compared to CompoDistill, which relies on uniform prompt-conditioned attention supervision, TRAG achieves stronger overall results by transitioning to a decomposed token-level objective targeting response-conditioned attention, e.g., improving Avg_8 from 61.9 to 62.9 on Qwen1.5-1.8B. This shift indicates that grounding cues manifested during decoding are particularly informative for distillation when supervised at token-level granularity. By leveraging this signal in the response phase, TRAG provides a more targeted supervision pathway, yielding robust improvements in downstream VQA across backbones.

Table 2: Compositional reasoning benchmark comparison of TRAG. Performance on three compositional reasoning benchmarks across Qwen1.5 and Qwen2.5 backbones at multiple parameter scales. Models are grouped by backbone and size. Within each backbone/size group, the best result is highlighted in **bold**. Blue indicates the teacher model for each backbone, gray rows represent models trained with the standard PT-SFT pipeline, yellow denotes prior distillation baselines, and green indicates our proposed method.

LLM	Method	Size	Compositional Reasoning Benchmarks			*Avg*
			SugarCrepe	BiVLC	Winoground	
Qwen1.5	TinyLLaVA	4B	87.3	93.2	70.1	83.5
	TinyLLaVA	1.8B	74.8	83.8	62.6	73.7
	KD		76.1	85.3	61.5	74.3
	LLaVA-KD		76.6	85.5	60.9	74.3
	CompoDistill		81.7	86.5	62.2	76.8
	TRAG		**83.1**	**89.1**	**67.2**	**79.8**
	TinyLLaVA	0.5B	52.8	56.5	50.2	53.2
	KD		51.0	50.1	52.1	51.0
	LLaVA-KD		66.9	74.5	51.7	64.4
	CompoDistill		72.1	78.6	56.6	69.1
	TRAG		**73.4**	**80.2**	**57.3**	**70.3**
Qwen2.5	TinyLLaVA	3B	88.1	94.6	75.0	85.9
	TinyLLaVA	1.5B	**87.9**	93.3	70.3	83.8
	TRAG		87.3	**93.8**	**73.9**	**85.0**
	TinyLLaVA	0.5B	75.1	85.1	57.0	72.4
	TRAG		**78.9**	**87.1**	**59.6**	**75.2**

Compositional Reasoning (CR) benchmarks. Table 2 reports reasoning performance on SugarCrepe, BiVLC, and Winoground. TRAG substantially strengthens compositional robustness, particularly in low-capacity regimes. For Qwen1.5-0.5B and 1.8B, TRAG improves the average scores to 70.3 and 79.8, exceeding the strongest prior baseline, CompoDistill, by +1.2 and +3.0 points, respectively. Similar gains are observed in the Qwen2.5 family, where the 1.5B student reaches an average of 85.0, narrowing the gap with the 3B teacher to 0.9 points. These results highlight that token-wise guidance during decoding enhances the student's ability to ground compositional judgments in the right visual evidence, leading to robust gains on compositional reasoning benchmarks.

Backbone Generalization. We further evaluate TRAG on the MobileLLaMA backbone, where it improves the 1.4B student over the PT-SFT baseline by +3.0 VQA Avg and +7.3 CR Avg. Detailed results are provided in Appendix F.

6.3 Analysis of Attention Fidelity

We evaluate attention fidelity on *VIGC-InstData* [35] using layer-wise teacher–student ($A_{R\to V}$) cosine sim-

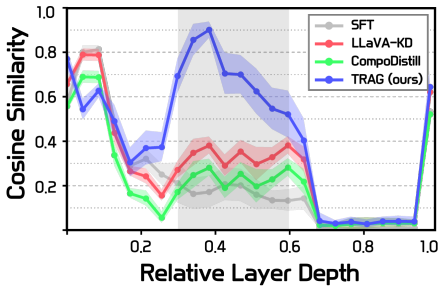

Fig. 5: Layer-wise Response-to-Vision attention fidelity.

ilarity during autoregressive generation (without teacher forcing), following the similarity definition in Section 4.1 (Figure 5). At the 30-60% layer depth, where cross-modal grounding is most intensive [7, 19], TRAG achieves an attention fidelity of approximately 0.72, significantly surpassing the SFT baseline (0.18) as well as LLaVA-KD (0.32) and CompoDistill (0.41). This elevated fidelity indicates that by supervising response-conditioned attention at the token level, TRAG successfully replicates the teacher's grounding logic throughout the generative process. Notably, this internal consistency directly correlates with the performance gains observed in Tables 1 and 2. These results confirm that maintaining high fidelity to the teacher's internal attention patterns is fundamental to achieving superior visual understanding and compositional robustness in low-capacity models.

6.4 Ablation Study

We report the main ablations on query-token selection and attention matching below. Additional ablations on relative-depth selection and response-supervision granularity are provided in Appendix F.

Query Token Type. Table 3 compares query token choices for forming visual-attention targets. Supervising only prompt tokens yields lower averages (VQA: 58.6, CR: 69.2) than using response tokens (VQA: 60.2, CR: 75.2). Using both prompt and response tokens does not improve VQA (58.4 vs. 58.6) and provides only a modest gain on CR (71.2 vs. 69.2), remaining substantially below response-only supervision. Overall, the best performance is achieved when supervision is concentrated on response tokens, suggesting that the most informative cross-modal signal emerges during response generation.

Table 3: Ablation results on query token types for attention extraction.

Query Token Type	VQA *Avg*	CR *Avg*
Prompt	58.6	69.2
Prompt + Response	58.4	71.2
Response (Ours)	**60.2**	**75.2**

Attention Matching Objectives. Table 4 compares alternative discrepancy measures for matching teacher–student response-to-vision attention. Among the fixed objectives, cosine similarity yields the strongest VQA average (58.7), while MSE performs best on CR (74.4), indicating that no single fixed measure dominates across benchmarks. The one-sided KL variants (forward or reverse) and the symmetric JSD objective also fall short of

Table 4: Ablation results on attention matching objectives.

Attention Matching	VQA *Avg*	CR *Avg*
MSE	58.2	74.4
Cosine Similarity	58.7	72.5
Forward KL	57.6	73.9
Reverse KL	58.1	73.4
JSD	57.9	71.6
Weighted Sum KL (Ours)	**60.2**	**75.2**

the best fixed baselines on at least one benchmark. In contrast, our entropy-aware weighted combination of forward and reverse KL achieves the best results on both VQA (60.2) and CR (75.2), improving over the strongest fixed baselines by +1.5 and +0.8, respectively. Overall, these gains suggest that effective attention distillation benefits from token-adaptive guidance rather than a globally fixed matching rule.

6.5 Scaling Experiments Across Model Sizes

Table 5 examines the impact of teacher capacity on distillation performance for a fixed 0.5B student model. Across both Qwen1.5 and Qwen2.5 backbones, TRAG consistently improves over the SFT baseline, with larger teachers yielding greater gains, especially on compositional reasoning. For Qwen1.5, CR Avg rises from 53.2 to 63.0 and 70.3 as the teacher scales from SFT to 1.8B and 4B; for Qwen2.5, it similarly increases from 72.4 to 74.3 and 75.2 as the teacher scales from SFT to 1.5B and 3B. These results indicate that TRAG effectively captures and transfers the enhanced grounding capabilities of larger teachers, demonstrating favorable scalability with respect to teacher model size.

Table 5: Scaling experiments with different teacher model sizes.

Teacher	Student	**VQA** *Avg*	**CR** *Avg*
Qwen1.5-0.5B (SFT)		55.8	53.2
Qwen1.5-1.8B	Qwen1.5-0.5B	57.3	63.0
Qwen1.5-4B		58.5	70.3
Qwen2.5-0.5B (SFT)		57.1	72.4
Qwen2.5-1.5B	Qwen2.5-0.5B	58.3	74.3
Qwen2.5-3B		60.2	75.2

7 Conclusion

We presented TRAG, a response-token-wise attention guidance objective for distilling MLLMs. By shifting cross-modal supervision from Prompt-to-Vision to Response-to-Vision attention, TRAG directly targets visual grounding during generation. It further decomposes supervision into token-level objectives and adaptively balances forward and reverse KL divergences according to the teacher's attention entropy. Integrated into a standard MLLM distillation pipeline, TRAG consistently improves performance on general VQA and compositional reasoning benchmarks while also improving teacher-student attention fidelity.

Acknowledgements

This work was supported by Samsung Electronics Co., Ltd (IO251223-14934-01) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics, and No.RS-2021-II211381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments).

References

1. Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Garea, S.R., Geist, M., Bachem, O.: On-policy distillation of language models: Learning from self-generated mistakes. In: The Twelfth International Conference on Learning Representations (2024)
2. Amara, I., Sepahvand, N., Meyer, B.H., Gross, W.J., Clark, J.J.: Bd-kd: Balancing the divergences for online knowledge distillation. arXiv preprint arXiv:2212.12965 (2022)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025)
4. Cai, Y., Zhang, J., He, H., He, X., Tong, A., Gan, Z., Wang, C., Xue, Z., Liu, Y., Bai, X.: Llava-kd: A framework of distilling multimodal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 239–249 (2025)
5. Cao, J., Zhang, Y., Huang, T., Lu, M., Zhang, Q., An, R., Ma, N., Zhang, S.: Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 19846–19856 (June 2025)
6. Chen, L., Zhao, X., Ding, K., Feng, W., Miao, C., Wang, Z., Guo, W., Wang, Y., Zheng, K., Zhang, B., Li, Z., Xiang, S.: Beyond next-token alignment: Distilling multimodal large language models via token interactions. arXiv preprint arXiv:2602.09483 (2026)
7. Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J.C., Geva, M., He, J., Wu, J., Li, M.: Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. arXiv preprint arXiv:2503.01773 (2025)
8. Chen, Y., Wang, P., Qin, G., Wu, W., Chen, M., Hao, Y.: Attention re-alignment in multimodal large language models via intermediate-layer guidance. Scientific Reports (2026)
9. Fan, Y., Tong, J., Zhao, A., Shen, X.: What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11987–11997 (2026)
10. Feng, Q., Li, W., Lin, T., Chen, X.: Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4178–4188 (2025)
11. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
12. Gu, Y., Dong, L., Wei, F., Huang, M.: MiniLLM: Knowledge distillation of large language models. In: The Twelfth International Conference on Learning Representations (2024)

13. Guo, Z., Zhang, P., Liang, P.: Sakd: Sparse attention knowledge distillation. Image and Vision Computing **146**, 105020 (2024)
14. Guo, Z., Yan, H., Li, H., Lin, X.: Class attention transfer based knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11868–11877 (2023)
15. Hendrycks, D.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
17. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in neural information processing systems **36**, 31096–31116 (2023)
18. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
19. Kim, J., Kim, K., Seo, S., Park, C.: Compodistill: Attention distillation for compositional reasoning in multimodal LLMs. In: The Fourteenth International Conference on Learning Representations (2026)
20. Ko, J., Kim, S., Chen, T., Yun, S.Y.: DistiLLM: Towards streamlined distillation for large language models. In: Forty-first International Conference on Machine Learning (2024)
21. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)
22. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
25. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
26. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
27. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
28. Miranda, I., Salaberria, A., Agirre, E., Azkune, G.: Bivlc: Extending vision-language compositionality evaluation with text-to-image retrieval. Advances in Neural Information Processing Systems **37**, 101880–101904 (2024)
29. Neo, C., Ong, L., Torr, P., Geva, M., Krueger, D., Barez, F.: Towards interpreting visual information processing in vision-language models. In: International Conference on Learning Representations. vol. 2025, pp. 57172–57189 (2025)
30. Pandey, R., Shao, R., Liang, P.P., Salakhutdinov, R., Morency, L.P.: Cross-modal attention congruence regularization for vision-language relation alignment. In: Pro-

ceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5444–5455 (2023)
31. Shu, F., Liao, Y., Zhang, L., Zhuo, L., Xu, C., Zhang, G., Shi, H., Chan, L., Yu, Z., He, W., et al.: Llava-mod: Making llava tiny via moe-knowledge distillation. In: The Thirteenth International Conference on Learning Representations (2025)
32. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
33. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
35. Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al.: Vigc: Visual instruction generation and correction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5309–5317 (2024)
36. Wang, C., Wang, Z., Xu, X., Tang, Y., Zhou, J., Lu, J.: Q-vlm: Post-training quantization for large vision-language models. Advances in Neural Information Processing Systems **37**, 114553–114573 (2024)
37. Wu, T., Tao, C., Wang, J., Yang, R., Zhao, Z., Wong, N.: Rethinking kullback-leibler divergence in knowledge distillation for large language models. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 5737–5755 (2025)
38. Xu, S., Li, X., Yuan, H., Qi, L., Tong, Y., Yang, M.H.: Llavadi: What matters for multimodal large language models distillation. arXiv preprint arXiv:2407.19409 (2024)
39. Yang, G., Yu, S., Sheng, Y., Yang, H.: Attention and feature transfer based knowledge distillation. Scientific Reports **13**(1), 18369 (2023)
40. Yoon, H.S., Yoon, E., Jang, J., Eom, S., Hong, J.W., Hasegawa-Johnson, M.A., Dai, Q., Luo, C., Yoo, C.D.: Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. In: Forty-third International Conference on Machine Learning (2026)
41. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
42. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017)
43. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
44. Zhang, Q., Cheng, A., Lu, M., Zhang, R., Zhuo, Z., Cao, J., Guo, S., She, Q., Zhang, S.: Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20857–20867 (2025)