

# Distilling Multimodal Evolution with Adaptive Anchors for Cloud-Edge Collaborative MLLMs

Chunyun Zhang<sup>1,2</sup>, Qilong Song<sup>1,2</sup>, Chaoran Cui<sup>1,2\*</sup>, Shuai Gong<sup>1,2</sup>,  
Fujun Yang<sup>1,2</sup>, Wenna Wang<sup>1,2</sup>, Peiguang Lin<sup>1,2</sup>

<sup>1</sup>School of Computing and Artificial Intelligence,  
Shandong University of Finance and Economics

<sup>2</sup>Shandong Key Laboratory of

Lightweight Intelligent Computing and Visualization for Digital Economy

Correspondence: crcui@sdufe.edu.cn

## Abstract

Cloud-edge collaboration provides a practical paradigm for deploying MLLMs on resource-constrained devices, where compact edge-side models are improved with supervision from stronger cloud-side models. However, existing cloud-edge distillation methods mainly rely on pseudo labels or static feature alignment, overlooking how the teacher progressively transforms multimodal inputs into answers. In this paper, we propose A<sup>2</sup>-MED, Adaptive Anchor-Guided Multimodal Evolution Distillation for cloud-edge MLLMs. Our key idea is to organize the teacher’s layer-wise multimodal evolution into functional stages through adaptive teacher-side anchoring, and transfer the resulting stage-wise evolution as process-level supervision. Specifically, we adaptively identify functional anchors from the teacher according to layer-wise indicators of cross-modal binding, semantic integration, and answer prediction, and then project these anchors to the student to establish corresponding functional stages. Under these teacher-defined stages, we align teacher-student anchor states and distill inter-stage multimodal evolution relations, enabling the student to learn how the teacher progressively transforms multimodal inputs into answers. Experiments on A-OKVQA and VizWiz show that our method consistently improves over representative cloud-edge and distillation baselines, with further analyses validating the effectiveness of adaptive anchors and evolution distillation.

## 1 Introduction

Multimodal large language models (MLLMs) have achieved strong performance in vision-language understanding and reasoning by integrating visual encoders with large language models (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Zhu et al., 2024; Bai et al., 2023). However, their rapidly

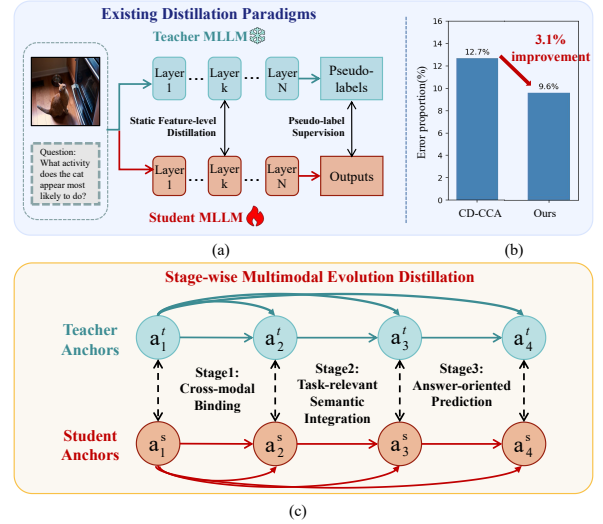


Figure 1: Motivation of our A<sup>2</sup>-MED. (a) Existing methods mainly transfer pseudo labels or static features. (b) CD-CCA leaves 12.7% teacher-correct but student-wrong samples, while our method reduces this ratio to 9.6%. (c) Our method distills the teacher’s stage-wise multimodal evolution, guiding the student to learn how the final answer is formed.

growing scale (Wen et al., 2025) makes direct deployment on edge devices impractical, as inference on such devices is often constrained by limited computation, memory, and communication budgets. Cloud-edge collaboration therefore offers a practical deployment paradigm, in which a compact edge-side MLLM performs local inference while a stronger cloud-side MLLM provides supervision for adaptation and performance improvement (Zhou et al., 2019; Wang et al., 2024).

Existing cloud-edge collaborative methods, such as CD-CCA (Wang et al., 2024), have shown that a powerful cloud-side MLLM can improve a compact edge-side MLLM by providing external supervision. Nevertheless, as illustrated in Figure 1(a), these methods mainly transfer teacher knowledge through pseudo labels or static feature-level distillation, offering only limited guidance on how

\*Corresponding Author

the teacher transforms multimodal inputs into answers. This limitation is reflected in Figure 1(b): under CD-CCA, 12.7% of the samples correctly answered by the teacher are still mispredicted by the student. Such cases suggest that supervision based primarily on teacher outputs or static feature alignment cannot fully transfer the teacher’s multimodal reasoning process. As shown in Figure 1(c), we therefore argue that the student should learn not only the teacher’s final answer or isolated feature representations, but also its stage-wise multimodal evolution, namely how visual and textual evidence is progressively transformed into answer-oriented representations.

For MLLMs, final answers are produced through a progressive layer-wise multimodal evolution process rather than from a single representation. Recent analyses suggest that shallow layers mainly capture visual grounding and cross-modal interaction, middle layers aggregate task-relevant multimodal semantics, and deeper layers are increasingly tied to answer prediction (Zhang et al., 2025; Fan et al., 2025). Such evolution can be interpreted as a coarse progression across functional stages, moving from cross-modal binding to task-relevant semantic integration and finally to answer-oriented prediction. Compared with supervision based only on final answers or static feature alignment, stage-wise evolution provides richer process-level knowledge. It is particularly valuable for compact edge-side students, whose limited capacity makes them more prone to incomplete cross-modal binding and insufficient semantic integration under distribution shift (Cui et al., 2025a,b).

However, transferring such stage-wise multimodal evolution is non-trivial, because effective inter-stage distillation requires the teacher and student to operate on corresponding functional stages. A straightforward strategy is to align fixed layers or uniformly divided layer groups. Yet this strategy implicitly assumes that different MLLMs follow the same functional progression across depth, which is often unrealistic in cloud-edge scenarios. The cloud-side teacher is typically deeper and more capable (Zhang et al., 2024; Jin et al., 2025), whereas the edge-side student is smaller and may compress, delay, or even fail to develop certain representation transitions. Consequently, fixed layer matching may align hidden states that correspond to different stages of multimodal evolution.

Another possible strategy is to independently discover anchors for the teacher and the student using

the same criteria. However, this may still produce misaligned functional stages, because the student’s anchor signals can be dominated by its underdeveloped representation dynamics rather than by the teacher’s desired evolution structure. We therefore use the teacher model as the reference for stage discovery: key transition anchors are adaptively identified from the teacher’s own layer-wise multimodal evolution and then projected to the student through proportional mapping. In this way, the teacher defines semantically grounded stage boundaries, while the student is guided to align with the corresponding functional stages under its compact depth and capacity constraints.

To this end, we propose A<sup>2</sup>-MED, Adaptive Anchor-Guided Multimodal Evolution Distillation for cloud-edge MLLMs. In our workflow, the edge-side student uploads high-entropy samples to the cloud, where a frozen teacher and a trainable student copy perform multimodal evolution distillation; the updated lightweight student parameters are then sent back to the edge device. Our key idea is to organize the teacher’s layer-wise multimodal evolution into functional stages through adaptive teacher-side anchoring, and transfer the resulting stage-wise evolution as process-level supervision. Specifically, we adaptively identify functional anchors from the teacher according to layer-wise indicators of cross-modal binding, semantic integration, and answer prediction, and then project these anchors to the student to establish corresponding functional stages. Under these teacher-defined stages, we align teacher-student anchor states and distill inter-stage multimodal evolution relations, enabling the student to learn how the teacher progressively transforms multimodal inputs into answers.

Our contributions are summarized as follows:

- We propose an adaptive anchor-guided multimodal evolution distillation framework for cloud-edge MLLMs, which transfers the teacher’s stage-wise multimodal evolution as process-level supervision beyond pseudo labels and static features.
- We introduce a teacher-guided adaptive anchor selection strategy that identifies functional stage anchors from the cloud-side teacher and maps them to the student, thereby establishing semantically matched stages for anchor state alignment and inter-stage multimodal evolution distillation.

- Experiments on A-OKVQA and VizWiz show that our method consistently outperforms strong baselines under cloud-edge adaptation settings, and further analyses confirm the effectiveness of the proposed components.

## 2 Related work

**MLLMs.** Multimodal large language models (MLLMs) have shown strong capabilities in visual question answering, image captioning, and cross-modal reasoning by coupling visual encoders with large language models. Representative works, including Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVA (Liu et al., 2023), demonstrate the effectiveness of large-scale visual instruction tuning. However, their large parameter sizes hinder deployment on resource-constrained edge devices. Recent studies therefore explore efficient MLLMs through token reduction, pruning, and lightweight adaptation, such as token-adaptive layer-wise pruning in MLLMs (Lee et al., 2025). Nevertheless, these works mainly focus on reducing model cost, while the intermediate semantic evolution of multimodal reasoning remains insufficiently explored.

**Knowledge distillation.** Knowledge distillation (KD) compresses large models by transferring teacher knowledge to compact students, initially via softened output distributions (Hinton et al., 2015). Later methods further exploit hidden representations and structural relations, such as patient KD (Sun et al., 2019) and RKD (Park et al., 2019). In multimodal settings, existing methods mainly focus on output supervision, static feature alignment, or local relational transfer. In contrast, our method distills the teacher’s layer-wise multimodal evolution trajectory through evolution initialization alignment, inter-stage evolution distillation, and anchor-wise cross-modal alignment.

**Cloud-edge collaborative learning.** Cloud-edge collaboration supports efficient deep model deployment by distributing computation between edge devices and cloud servers to balance latency, bandwidth, and energy cost (Zhou et al., 2019; Kang et al., 2017; Li et al., 2018). Early methods such as Neurosurgeon and Edgent focus on collaborative inference through partitioning and adaptive offloading (Kang et al., 2017; Li et al., 2018). Recent works extend this paradigm to multimodal model adaptation and distillation, such as CD-CCA

(Wang et al., 2024). However, most methods treat the cloud model as a pseudo label or feature supervisor, whereas our work transfers stage-aware multimodal evolution trajectories to improve edge-side reasoning.

## 3 Methods

We propose **A<sup>2</sup>-MED**, an Adaptive Anchor-guided Multimodal Evolution Distillation framework for cloud-edge collaborative MLLMs. The overall framework is illustrated in Figure 2.

### 3.1 Overall Framework of A<sup>2</sup>-MED

A<sup>2</sup>-MED transfers the teacher’s stage-wise multimodal evolution to an edge-side student under cloud-edge collaboration. Instead of using the cloud teacher only for pseudo labels or static feature supervision, it defines functional anchors along the teacher’s layer-wise evolution and uses them as stage coordinates for student distillation.

As shown in Figure 2, the edge-side student performs local inference and uploads high-entropy samples to the cloud, where a frozen teacher and a trainable student copy are maintained. For uploaded samples, the teacher identifies functional anchors using three teacher-side indicators, which are then proportionally mapped to the student. Based on the aligned stages, A<sup>2</sup>-MED aligns anchor states, distills inter-stage multimodal evolution relations, and sends the updated lightweight student back to the edge for subsequent inference and sample selection.

### 3.2 Edge-side Sample Uploading

In cloud-edge collaborative MLLMs, uploading all edge-side samples to the cloud is costly and often unnecessary. We therefore upload samples that are likely to expose the student’s current understanding gaps, using predictive entropy as the selection criterion.

For an input sample  $x$ , the edge-side model  $f_s$  performs forward inference to obtain the probability distribution  $P_s(x) \in R^{|v|}$  of the model output, and its entropy is computed as:

$$H(P_s(x)) = - \sum_{v \in V} P_s(v|x) \log P_s(v|x). \quad (1)$$

We upload samples whose entropy is larger than a threshold  $\tau$ :  $X^* = \{x \mid H(P_s(x)) > \tau\}$ . The selected samples are then used for cloud-side multimodal evolution distillation.

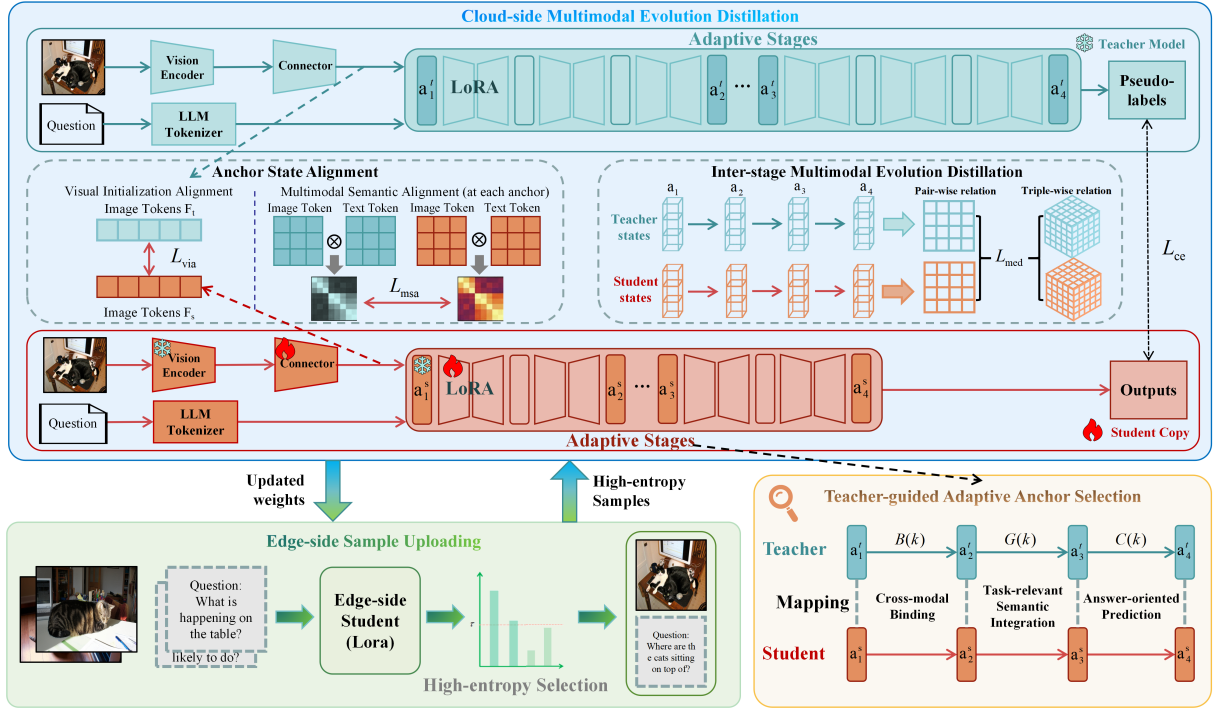


Figure 2: Overall framework of A<sup>2</sup>-MED. The edge-side student uploads high-entropy samples that are likely to expose its current understanding gaps. On the cloud side, a frozen teacher MLLM defines functional stage anchors and maps them to the trainable student copy, enabling distillation on semantically corresponding stages. A<sup>2</sup>-MED then aligns anchor states and transfers inter-stage multimodal evolution relations, after which the updated lightweight student parameters are sent back to the edge device.

### 3.3 Teacher-guided Adaptive Anchor Selection

After selected samples are uploaded to the cloud, A<sup>2</sup>-MED first aligns teacher and student stages semantically. Rather than matching fixed layers, we use the cloud-side teacher as the reference to define functional anchors, since it provides the multimodal evolution trajectory to transfer.

Specifically, we fix the first and last layers as boundaries and adaptively search two intermediate teacher anchors. These anchors partition the teacher’s layer-wise evolution into three stages: cross-modal binding, task-relevant semantic integration, and answer-oriented prediction. The teacher anchors are then mapped to the student by relative layer positions, yielding aligned stage boundaries for subsequent distillation.

To locate functional anchors, we define three teacher-side indicators that characterize the progress of different stages. For cross-modal binding, we measure how strongly text tokens are associated with relevant visual evidence. Given the visual tokens  $\{v_i^k\}_{i=1}^{N_v}$  and text tokens  $\{t_j^k\}_{j=1}^{N_t}$  at teacher layer  $k$ , where  $N_v$  and  $N_t$  denote the numbers of visual and text tokens, respectively, the

binding strength is defined as:

$$B(k) = \frac{1}{N_t} \sum_{j=1}^{N_t} \max_{1 \leq i \leq N_v} \text{sim}(v_i^k, t_j^k). \quad (2)$$

For task-relevant semantic integration, we measure how the question-related representation interacts with both visual and textual tokens. Let  $q^k$  denote the question-related representation at teacher layer  $k$ . The integration strength is defined as:

$$s_i^{v,k} = \text{sim}(q^k, v_i^k), \quad s_j^{t,k} = \text{sim}(q^k, t_j^k), \quad (3)$$

$$G(k) = \frac{1}{N_v} \sum_i s_i^{v,k} + \frac{1}{N_t} \sum_j s_j^{t,k}. \quad (4)$$

For answer-oriented prediction, we measure how well the representation at layer  $k$  supports the teacher-generated answer. Let  $h^k$  denote the hidden state at the answer-prediction position. By feeding  $h^k$  into the teacher’s original language modeling head, the answer predictability is computed as:

$$C(k) = p(\hat{y} | h^k). \quad (5)$$

Based on these indicators, we select anchors by maximizing the stage-wise progress of the teacher’s

multimodal evolution. Let  $a_1$  and  $a_4$  denote the first and last anchors, and let  $a_2$  and  $a_3$  be two candidate intermediate anchors. The progress of the three stages is defined as:

$$S_{\text{bind}}(a_1, a_2) = B(a_2) - B(a_1), \quad (6)$$

$$S_{\text{reason}}(a_2, a_3) = G(a_3) - G(a_2), \quad (7)$$

$$S_{\text{answer}}(a_3, a_4) = C(a_4) - C(a_3). \quad (8)$$

The optimal intermediate anchors are obtained by maximizing the weighted sum of these stage-wise scores:

$$S_{\text{total}} = \lambda_b S_{\text{bind}} + \lambda_r S_{\text{reason}} + \lambda_a S_{\text{answer}}, \quad (9)$$

$$(a_2^*, a_3^*) = \arg \max_{\substack{a_2 < a_3, \\ a_2 \in [1, \lfloor \rho L_t \rfloor], \\ a_3 \in [\lceil \rho L_t \rceil, L_t - 2], \\ a_3 - a_2 \geq \delta}} S_{\text{total}}. \quad (10)$$

After obtaining the teacher anchors, we project them to the student according to their relative layer positions:

$$a_i^s = \text{round} \left( \frac{a_i^t}{L_t - 1} \times (L_s - 1) \right). \quad (11)$$

In this way, the teacher defines a stage-wise evolution coordinate system, while the student obtains corresponding anchor positions under its own depth. These anchors provide the basis for the following anchor state alignment and inter-stage multimodal evolution alignment.

### 3.4 Anchor State Alignment

After adaptive anchor selection, A<sup>2</sup>-MED aligns the corresponding teacher-student anchor states before distilling inter-stage evolution. This is achieved through Visual Initialization Alignment and Anchor-wise Multimodal Semantic Alignment.

**Visual Initialization Alignment (VIA)** aligns the initial image-token representations after the visual projection layer, providing a stable visual starting point for multimodal evolution distillation. Let  $F_s^{\text{img}} \in \mathbb{R}^{B \times L \times C_s}$  and  $F_t^{\text{img}} \in \mathbb{R}^{B \times L \times C_t}$  denote the teacher and student image-token representations after the visual projection layer, respectively, where  $L$  is the number of image tokens after the projection layer. Since the teacher and student may have different hidden dimensions, we project the student features into the teacher feature space with a learnable  $1 \times 1$  convolution  $\phi_{\text{img}}$  and minimize their distance:

$$\mathcal{L}_{\text{via}} = \frac{1}{B \cdot L \cdot C_t} \left\| \phi_{\text{img}}(F_s^{\text{img}}) - F_t^{\text{img}} \right\|_F^2. \quad (12)$$

### Anchor-wise Multimodal Semantic Alignment (MSA)

aligns the multimodal interaction patterns at each corresponding anchor. For the  $i$ -th anchor pair, we extract image-token and text-token features from the teacher and student hidden states, denoted as  $(V_t^i, T_t^i)$  and  $(V_s^i, T_s^i)$ , respectively. After projecting the student features into the teacher feature space, we compute their cross-modal correlation matrices:

$$M_t^i = \text{norm}(V_t^i) \cdot \text{norm}(T_t^i)^\top, \quad (13)$$

$$M_s^i = \text{norm}(V_s^i) \cdot \text{norm}(T_s^i)^\top. \quad (14)$$

The multimodal semantic alignment loss is defined as:

$$\mathcal{L}_{\text{msa}} = \frac{1}{N} \sum_{i=1}^N \|M_s^i - M_t^i\|_F^2. \quad (15)$$

### 3.5 Inter-stage Multimodal Evolution Distillation

After establishing stable anchor-wise multimodal states, Multimodal Evolution Distillation (MED) distills the teacher's evolution trajectory to transfer structural transitions across stages. This enables the student to learn how the teacher progresses from visual grounding to semantic reasoning and answer prediction.

Specifically, we build inter-layer relations over key layers selected by adaptive anchors. Inspired by CKD (Park et al., 2021), which introduces Layer Transforming Relation (LTR) to model structural knowledge via pair-wise cosine and triple-wise angular relations, we extend this idea to the full multimodal sequence by concatenating image and text tokens. Thus, the evolution of visual tokens from shallow representations to deep semantics is also constrained.

For the hidden layer representation of the  $k$ -th anchor layer, take a complete sequence slice concatenated from the image and text:  $\tilde{h}_k = h_k[:, p_{\text{img}} : p_{\text{end}}, :]$ . In which,  $p_{\text{img}}$  denotes the starting position of the image tokens, and  $p_{\text{end}}$  denotes the effective ending position of the text tokens.

The pair-wise cosine relation captures the magnitude of feature variation between any two anchor layers. For any  $k < m$ , we define it as:

$$\Phi(k, m; h) = \frac{\tilde{h}_k \cdot \tilde{h}_m}{\|\tilde{h}_k\|_2 \cdot \|\tilde{h}_m\|_2}, \quad (16)$$

$$\mathcal{L}_{pair} = \frac{1}{\binom{N}{2}} \sum_{0 \leq k < m \leq N-1} \|\Phi(k, m; h_s) - \Phi(k, m; h_t)\|_F^2. \quad (17)$$

The triple-wise angular relation takes the intermediate layer as the vertex and captures a higher-order angular structure of the evolution direction among three layers. For any  $k < m < o$ , we define it as:

$$\Psi(k, m, o; h) = \frac{(\tilde{h}_k - \tilde{h}_m)^\top (\tilde{h}_o - \tilde{h}_m)}{\|\tilde{h}_k - \tilde{h}_m\|_2 \|\tilde{h}_o - \tilde{h}_m\|_2}, \quad (18)$$

$$\mathcal{L}_{triple} = \frac{1}{\binom{N}{3}} \sum_{0 \leq k < m < o \leq N-1} \|\Psi(k, m, o; h_s) - \Psi(k, m, o; h_t)\|_F^2. \quad (19)$$

By combining the two, the MED distillation loss is defined as:

$$\mathcal{L}_{med} = \mathcal{L}_{pair} + \lambda_{triple} \mathcal{L}_{triple}. \quad (20)$$

### 3.6 Overall Objective

To simultaneously preserve the task modeling capability of the student model and facilitate the transfer of evolution trajectories, we jointly optimize the language modeling loss together with the proposed trajectory distillation losses. Specifically, the teacher model generates pseudo labels for each sample, and the student model computes the language modeling loss based on the pseudo labels and its own predictions:

$$\mathcal{L}_{ce} = -\frac{1}{|X^*|} \sum_{x \in X^*} \log p_s(\hat{y} | x). \quad (21)$$

By combining the language modeling loss with the aforementioned distillation losses, the overall objective function is formulated as:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{via} + \beta \mathcal{L}_{msa} + \gamma \mathcal{L}_{med}, \quad (22)$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are balancing coefficients for the loss terms. After cloud-side adaptation on the current batch, the updated weights are sent back to the edge side, which then uses them to select the next batch of samples.

## 4 Experiments

In this section, we detail the experimental setup and report a series of results to validate the effectiveness of the proposed method.

| Method                | Accuracy      | Gain         |
|-----------------------|---------------|--------------|
| Source-only           | 56.86%        | /            |
| TENT-continual        | 56.59%        | -0.27%       |
| PKD                   | 58.18%        | 1.32%        |
| ChannelWiseDivergence | 57.24%        | 0.38%        |
| CD-CCA                | 58.57%        | 1.71%        |
| <b>Ours</b>           | <b>60.49%</b> | <b>3.63%</b> |

Table 1: A comparison of A-OKVQA continuous adaptability performance.

| Method                | Accuracy      | Gain         |
|-----------------------|---------------|--------------|
| Source-only           | 31.69%        | /            |
| TENT-continual        | 12.60%        | -19.09%      |
| PKD                   | 32.97%        | 1.28%        |
| ChannelWiseDivergence | 32.46%        | 0.77%        |
| CD-CCA                | 33.45%        | 1.76%        |
| <b>Ours</b>           | <b>34.52%</b> | <b>2.83%</b> |

Table 2: A comparison of VizWiz continuous adaptability performance.

### 4.1 Experimental Setup

**Datasets and Metrics.** We evaluate our method on A-OKVQA and VizWiz. A-OKVQA (Schwenk et al., 2022) is an open-domain VQA benchmark requiring both visual understanding and common-sense knowledge, while VizWiz (Gurari et al., 2018) contains real-world images captured by visually impaired users and often involves blur, poor lighting, occlusion, and irregular composition. These datasets cover both general vision-language understanding and challenging low-quality visual inputs. We report the official VQA accuracy (Schwenk et al., 2022) on both datasets.

**Baselines.** We compare A<sup>2</sup>-MED with five representative baselines: Source-only (Gao et al., 2023), TENT-continual (Wang et al., 2020), PKD (Cao et al., 2022), ChannelWiseDivergence (Shu et al., 2021), and CD-CCA (Wang et al., 2024). These methods cover no-adaptation, test-time adaptation, feature-level distillation, and cloud-edge collaborative distillation settings.

**Implementation Details.** The cloud-side teacher and edge-side student are implemented with LLaMA-Adapter (Gao et al., 2023) built on LLaMA2-13B (Touvron et al., 2023) and LLaMA2-7B (Touvron et al., 2023), respectively, and both models use BLIP-2 (Li et al., 2023) as the vi-

sual encoder. Both models are initialized from source models instruction-tuned on GPT4LLM (Peng et al., 2023) instruction data and VQA v2 (Goyal et al., 2017), and then adapted to target-domain samples from A-OKVQA or VizWiz in a sequential cloud-edge setting. The teacher is kept frozen during adaptation, and only lightweight student parameters are updated. Detailed hyperparameters are provided in Appendix A.

## 4.2 Performance Comparison

Table 1 and Table 2 report the performance on A-OKVQA and VizWiz. A<sup>2</sup>-MED achieves the best performance on both datasets, showing the effectiveness of transferring the teacher’s stage-wise multimodal evolution in cloud-edge adaptation.

On A-OKVQA, A<sup>2</sup>-MED reaches 60.49% accuracy, outperforming the most related cloud-edge baseline, CD-CCA, by 1.92%. This gain suggests that modeling the teacher’s multimodal evolution provides additional supervision beyond the pseudo-label and feature-level guidance used in existing cloud-edge distillation methods. Although PKD and ChannelWiseDivergence improve over Source-only, they still lag behind A<sup>2</sup>-MED, suggesting that static feature-level distillation cannot fully capture the teacher’s progressive transformation from multimodal inputs to answers.

On VizWiz, A<sup>2</sup>-MED achieves 34.52% accuracy and surpasses CD-CCA by 1.07%. Since VizWiz contains low-quality real-world images with blur, poor lighting, and irregular composition, this result shows that our method remains effective under more challenging visual conditions. In contrast, TENT-continual suffers a severe performance drop, suggesting that directly minimizing prediction entropy can be unstable when the model faces highly noisy visual inputs. Overall, these results show that adaptive anchor-guided multimodal evolution distillation provides useful process-level supervision beyond final predictions and static feature-level distillation.

## 4.3 Ablation Studies

To further examine the effectiveness of the key components in A<sup>2</sup>-MED, we conduct ablation experiments on VIA, MSA, and MED. As shown in Table 3, removing any component leads to a performance drop, indicating that these modules provide complementary supervision for multimodal evolution distillation.

| Component                        | Accuracy(%)   |
|----------------------------------|---------------|
| Full model (A <sup>2</sup> -MED) | <b>60.49</b>  |
| w/o VIA                          | 59.42 (-1.07) |
| w/o MSA                          | 59.30 (-1.19) |
| w/o MED                          | 59.71 (-0.78) |

Table 3: Ablation study of different components on A<sup>2</sup>-MED.

| $\tau$ | Num. Samples | Accuracy (%) |
|--------|--------------|--------------|
| 1      | 887          | 58.49        |
| 2      | 659          | 58.52        |
| 3      | <b>458</b>   | <b>60.49</b> |
| 4      | 329          | 57.87        |
| 5      | 216          | 59.13        |

Table 4: The impact of the entropy threshold  $\tau$  for sample selection.

Specifically, removing VIA decreases the accuracy from 60.49% to 59.42%, indicating that aligning initial visual representations is beneficial for subsequent multimodal distillation. Removing MSA leads to the largest drop, from 60.49% to 59.30%, suggesting that anchor-wise multimodal semantic alignment is important for capturing the teacher’s image-text interaction patterns at corresponding stages. When MED is removed, the accuracy drops to 59.71%, which shows that inter-stage multimodal evolution relations provide additional process-level supervision beyond anchor state alignment.

## 4.4 Analysis of Edge-side Sample Uploading

Table 4 analyzes the effect of the entropy threshold  $\tau$  on edge-side sample uploading. As  $\tau$  increases, the number of uploaded samples decreases, showing that the entropy threshold effectively controls the amount of samples used for cloud-side adaptation. A<sup>2</sup>-MED achieves the best performance at  $\tau = 3$ , where 458 samples are uploaded and the accuracy reaches 60.49%. This result suggests that moderately selecting high-entropy samples helps the cloud-side distillation focus on informative examples that better expose the student’s current understanding gaps. When  $\tau$  is too small, the selection criterion becomes overly loose, and many low-uncertainty samples are uploaded, diluting cloud-side distillation with redundant signals. When  $\tau$  is too large, the selected samples become too few to

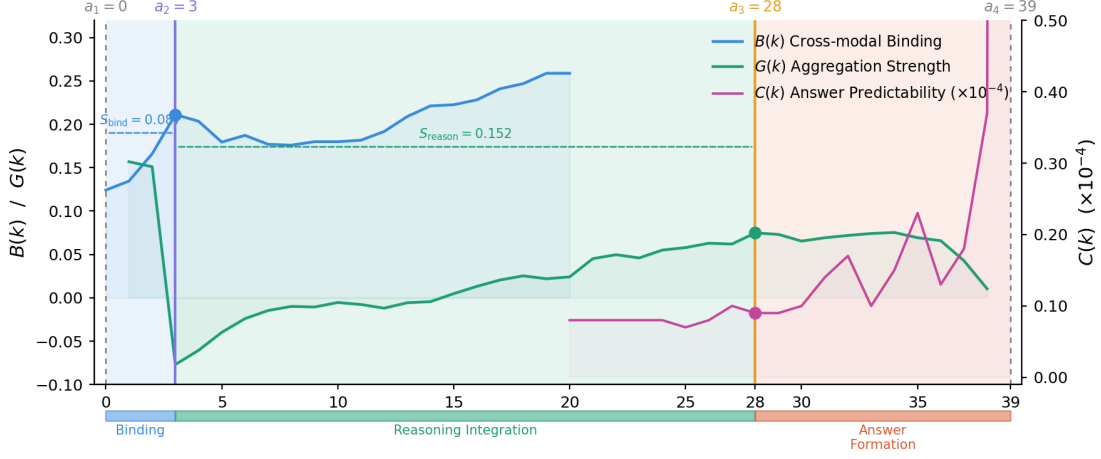


Figure 3: In the adaptive anchor selection strategy, the curves of the three indicators  $B(k), G(k), C(k)$  changing with the number of layers.

sufficiently cover the target-domain distribution. Overall, these results show that entropy-guided sample uploading provides a practical balance between informative sample selection and adaptation effectiveness in cloud-edge MLLMs.

#### 4.5 Anchor Selection Strategy Analysis

To examine the necessity of teacher-guided anchor selection, we compare our strategy with random anchors, uniformly spaced anchors, and independently selected anchors, where the teacher and student select anchors separately using the same criteria. As shown in Table 5, random anchors lead to a 2.12% accuracy drop, showing that arbitrary layer selection provides unstable supervision. Uniform anchors also reduce the accuracy by 1.28%, indicating that fixed layer partitioning cannot properly reflect the uneven functional progression of MLLMs across depth. More importantly, independently selected anchors cause a 1.89% drop, even though the same indicators are used for both models. This suggests that the teacher and student may not reach corresponding functional stages at the same relative positions, and separately selected anchors may still lead to mismatched stage alignment. These results support our design of discovering anchors from the teacher and then mapping them to the student for multimodal evolution distillation.

To further examine whether the teacher-side indicators capture the assumed functional stages, Figure 3 visualizes the three teacher-side indicators across layers.  $B(k)$  increases in shallow layers, suggesting that the teacher first establishes cross-modal binding between visual and textual tokens. As  $B(k)$  becomes stable,  $G(k)$  gradually increases,

| Strategy                       | Accuracy(%)  |
|--------------------------------|--------------|
| Random anchors                 | 58.37        |
| Uniform anchors                | 59.21        |
| Independently selected anchors | 58.60        |
| Adaptive anchors (ours)        | <b>60.49</b> |

Table 5: Comparison of different anchor selection strategies on A-OKVQA.

indicating a shift toward task-relevant multimodal semantic integration. In deeper layers,  $C(k)$  rises more clearly, showing that the representations become increasingly related to answer prediction. These trends are consistent with the functional stages assumed in our method and provide empirical support for using  $B(k)$ ,  $G(k)$ , and  $C(k)$  to identify teacher anchors for multimodal evolution distillation.

## 5 Conclusion

We presented A<sup>2</sup>-MED, an adaptive anchor-guided multimodal evolution distillation framework for cloud-edge MLLMs. A<sup>2</sup>-MED transfers the teacher’s stage-wise multimodal evolution as process-level supervision beyond pseudo labels and static feature-level distillation. By adaptively identifying functional anchors from the teacher and mapping them to the student, our method aligns corresponding anchor states and distills inter-stage evolution relations. Experiments on A-OKVQA and VizWiz show that A<sup>2</sup>-MED consistently outperforms representative baselines under cloud-edge adaptation settings.

## Limitations

Although A<sup>2</sup>-MED improves cloud-edge adaptation for MLLMs, it requires access to the intermediate representations of the cloud-side teacher for adaptive anchor selection and multimodal evolution distillation. This is feasible when the cloud model is controllable, but may not directly apply to black-box API-based teachers.

In addition, although entropy-guided sample uploading reduces unnecessary cloud communication, selected edge-side samples still need to be transmitted to the cloud, which may raise privacy and communication concerns in real-world deployment. Our experiments also focus on visual question answering benchmarks, and future work can further examine adaptive anchor-guided evolution distillation on broader multimodal tasks and privacy-preserving cloud-edge settings.

## Acknowledgement

This work was supported in part by the Young Scientists Fund (A Class) of Shandong Provincial Natural Science Foundation under Grant ZR2026QA03, the National Natural Science Foundation of China under Grant 62077033, and the Taishan Scholar Program of Shandong Province under Grant tsqn202211199.

## References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.
- Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. 2022. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35:15394–15406.
- Chaoran Cui, Ziyi Liu, Shuai Gong, Lei Zhu, Chunyun Zhang, and Hui Liu. 2025a. When adversarial training meets prompt tuning: Adversarial dual prompt tuning for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 34:1427–1440.
- Chaoran Cui, Yongrui Zhen, Shuai Gong, Chunyun Zhang, Hui Liu, and Yilong Yin. 2025b. Dynamic prompt allocation and tuning for continual test-time adaptation. *Science China Information Sciences*, 68(6):160109.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tjong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.
- Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. 2025. Visipruner: Decoding discontinuous cross-modal dynamics for efficient multimodal llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18896–18913.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and Lizhuang Ma. 2025. Efficient multimodal large language models: a survey. *Visual Intelligence*, 3(1):27.
- Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News*, 45(1):615–629.
- Jaewoo Lee, Keyang Xuan, Chanakya Ekbote, Sandeep Polisetty, Yi R Fung, and Paul Pu Liang. 2025.

- Tamp: Token-adaptive layerwise pruning in multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6892–6908.
- En Li, Zhi Zhou, and Xu Chen. 2018. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In *Proceedings of the 2018 workshop on mobile edge communications*, pages 31–36.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Geondo Park, Gyeongman Kim, and Eunho Yang. 2021. Distilling linguistic context for language model compression. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 364–378.
- Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.
- Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. 2021. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5311–5320.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4323–4332.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.
- Guanqun Wang, Jiaming Liu, Chenxuan Li, Yuan Zhang, Junpeng Ma, Xinyu Wei, Kevin Zhang, Maurice Chong, Renrui Zhang, Yijiang Liu, and Shanghang Zhang. 2024. Cloud-device collaborative learning for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12646–12655.
- Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15537–15549.
- Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430.
- Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19781–19791. IEEE.
- Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. 2019. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8):1738–1762.
- Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.

## A Implementation Details

The trainable parameters of the student model include the LoRA modules (Hu et al., 2022), bias terms, normalization-layer parameters, and the projection layer, while the teacher remains frozen throughout training. We optimize the model with AdamW (Loshchilov and Hutter, 2017) using a learning rate of  $3 \times 10^{-5}$ . The loss weights are set to  $\alpha = 0.5$ ,  $\beta = 100$ ,  $\gamma = 100$ , and  $\lambda_{\text{triple}} = 0.5$ ; the entropy threshold is 3. Before knowledge distillation, we fine-tune the model using 52K instruction data from GPT4LLM (Peng et al., 2023) and VQA v2 (Goyal et al., 2017).

## B Additional Experimental Results

### B.1 Token Range In Distillation

In the multimodal evolution distillation module, we extend the distillation target from text tokens alone to the multimodal sequence formed by concatenating image tokens and text tokens. To support this design, we conduct an ablation study to examine the performance differences introduced by different token selections. Specifically, Table 6 compares three settings: distillation applied only to text tokens, only to image tokens, and to the full multimodal sequence. The experimental results show that our design achieves the best performance. This is because multimodal large models rely on understanding visual information, and the major distribution discrepancy between the source and target domains also primarily arises from the visual modality. Restricting the distillation target to text tokens alone provides the student model with incomplete knowledge and fails to guide it toward robust visual understanding. On the other hand, limiting the target to image tokens discards the semantic guidance provided by text, thereby preventing the model from capturing how visual and linguistic features jointly evolve and interact across layers. Such holistic structural information cannot be provided by either modality alone.

| Token Range              | VQA Accuracy (%) |
|--------------------------|------------------|
| image-tokens             | 59.51            |
| text-tokens              | 58.25            |
| <b>all-tokens (ours)</b> | <b>60.49</b>     |

Table 6: Impact of token range on model performance.

| VIA | MSA | MED | Accuracy (%) |
|-----|-----|-----|--------------|
| ✓   |     |     | 59.21        |
| ✓   | ✓   |     | 59.71        |
| ✓   | ✓   | ✓   | <b>60.49</b> |

Table 7: Supplementary incremental ablation of the main components on A-OKVQA.

### B.2 Supplementary Incremental Ablation

To complement the main ablation study in Table 3, we provide an incremental analysis of the main components in Table 7. While the main ablation removes each component from the full model to examine its necessity, this supplementary analysis progressively enables VIA, MSA, and MED to show how the complete objective is built up.

Starting with VIA alone, the model achieves 59.21% accuracy, suggesting that visual initialization alignment provides a useful starting point for subsequent multimodal distillation. Adding MSA improves the accuracy to 59.71%, indicating that anchor-wise multimodal semantic alignment brings additional gains by aligning teacher-student image-text interaction patterns at corresponding stages. Finally, incorporating MED further improves the performance to 60.49%, showing that inter-stage evolution relations provide complementary process-level supervision after anchor states have been aligned. These results are consistent with the main ablation and provide additional evidence that the three objectives contribute cumulatively to A<sup>2</sup>-MED.

### B.3 Hyperparameter Sensitivity

To further study the influence of loss balancing coefficients, we analyze the performance trends of three key hyperparameters in Figure 4. Specifically,  $\alpha$  controls visual initialization alignment,  $\beta$  controls anchor-wise multimodal semantic alignment, and  $\gamma$  controls inter-stage multimodal evolution distillation.

As shown in Figure 4(a), increasing  $\alpha$  initially improves performance, indicating that aligning the initial visual representations between the teacher and student provides a stable foundation for subsequent multimodal distillation. However, an overly large  $\alpha$  degrades performance, since excessive emphasis on low-level visual initialization may restrict the student’s higher-level semantic modeling ability.

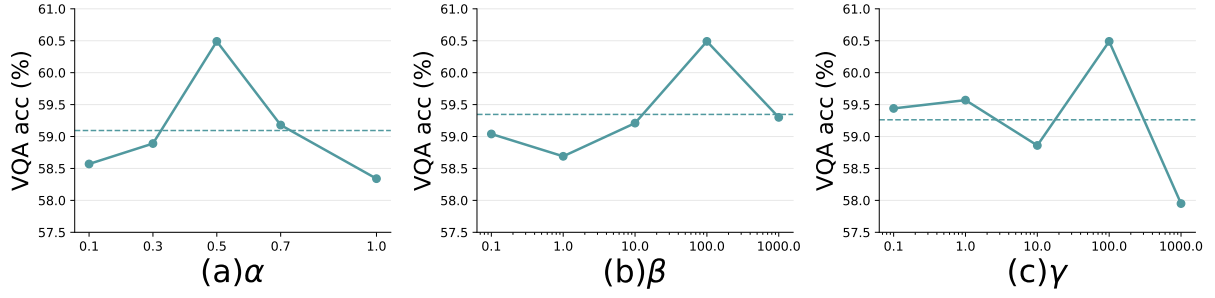


Figure 4: The impact of hyperparameters on A-OKVQA.

As shown in Figure 4(b),  $\beta$  also requires a balanced setting. When  $\beta$  is too small, the student cannot sufficiently inherit the teacher’s anchor-wise image-text interaction patterns; when it is too large, local cross-modal alignment may over-constrain the representation space and reduce the flexibility of multimodal reasoning.

As shown in Figure 4(c),  $\gamma$  determines the strength of MED, which transfers the teacher’s inter-stage evolution relations through pair-wise and triple-wise structural constraints. A small  $\gamma$  provides insufficient trajectory-level supervision, whereas a large  $\gamma$  may cause the structural distillation objective to dominate the language modeling objective. Overall, these results demonstrate that VIA, MSA, and MED provide complementary supervision, and properly balancing their weights is important for achieving the best performance.

#### B.4 Case Study On Teacher-Student Gap

To further analyze the remaining gap between the cloud-side teacher and the edge-side student, we present qualitative cases where the teacher predicts the correct answer while the baseline student fails. As shown in Figure 5, CD-CCA still struggles to faithfully transfer the teacher’s multimodal reasoning behavior in scenarios that require fine-grained visual grounding or context-aware inference. In contrast, our method produces answers that are more consistent with the teacher by distilling the teacher’s stage-wise multimodal evolution.

In the first case, CD-CCA predicts “red”, seemingly distracted by the salient color of the flower. Our method instead focuses on the center region and correctly answers “white”, consistent with the teacher’s prediction. This suggests that our method helps the student identify the visual evidence that is truly relevant to the textual query, rather than relying on dominant but misleading visual cues. In the second case, CD-CCA gives a generic dog-related

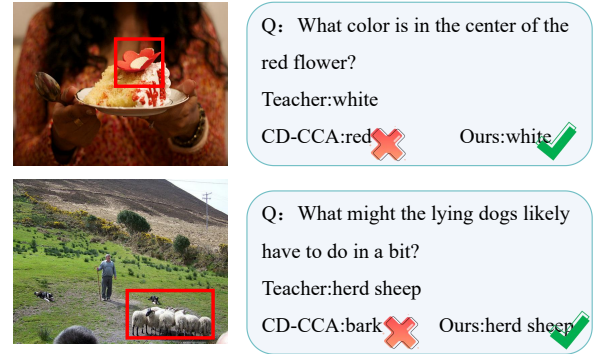


Figure 5: Qualitative comparison of teacher and student responses.

answer “bark”. By contrast, our method captures the surrounding sheep and infers the more contextually appropriate answer “herd sheep”. These examples indicate that our method improves not only final answer imitation, but also the student’s ability to organize relevant multimodal evidence in a teacher-consistent manner.