CROSS-MODAL FEATURE DYNAMICS DISTILLATION FOR RESOURCE-CONSTRAINED MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) achieve strong performance on vision–language tasks. However, their substantial computational cost limits deployment in resource-constrained settings. Knowledge distillation provides a promising route for transferring cross-modal reasoning knowledge from a large teacher to a compact student. Existing MLLM distillation methods, however, predominantly match static outputs or representations at isolated layers, overlooking the layer-wise evolutionary nature through which cross-modal reasoning is progressively formed. We introduce Cross-Modal Feature Dynamics Distillation (CM-FDD), which distills cross-modal grounding from two complementary perspectives: the grounding state at each depth and its evolution across selected depths. To enable this transfer across heterogeneous teacher–student models, we represent grounding with text-to-vision affinity, providing a model-comparable space that captures both where visual evidence is grounded and how that grounding changes over depth. When teacher and student use different visual resolutions, we further align their visual supports through deterministic cross-resolution projection. Experiments on ten benchmarks show that CM-FDD consistently outperforms prior MLLM distillation methods and achieves state-of-the-art performance under matched model-size settings, demonstrating the effectiveness of preserving grounding dynamics for efficient multimodal models.

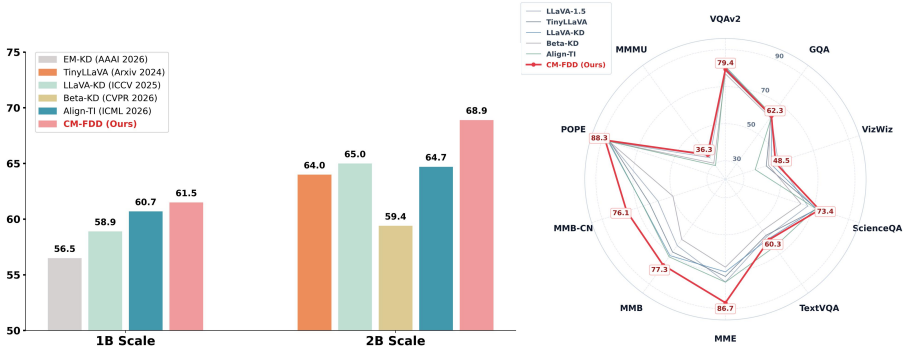


Figure 1: Left: Average performance across ten benchmarks at the 1B and 2B scales. Right: Per-benchmark comparison. CM-FDD achieves state-of-the-art performance across both model scales and a wide range of multimodal tasks.

1 INTRODUCTION

Multimodal large language models (Bai et al., 2023; Liu et al., 2023; Lin et al., 2024) have achieved impressive performance across a wide range of vision-language tasks, ranging from visual question answering to nuanced real-world reasoning and autonomous system interaction. But their substantial computational and memory costs hinder efficient deployment (Yu et al., 2024; Zhang et al., 2024; Xia et al., 2023). Knowledge distillation (Hinton et al., 2015) provides a practical approach to transferring the capabilities of large teacher MLLMs to compact students.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

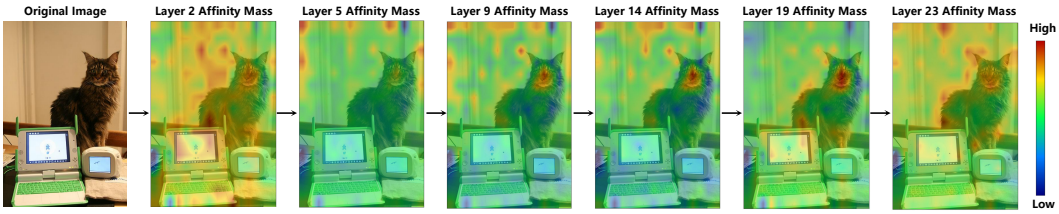


Figure 2: Evolution of cross-modal visual reasoning under the CM-FDD framework: visualization of the student model's intermediate-layer affinity mass allocation, showing how the token "cat" increasingly focuses on vision tokens spatially correlated with the cat across layers (from Layer 2 to Layer 23).

Recent studies (Cai et al., 2025; Feng et al., 2025; Chen et al., 2026; Kim et al., 2026; Shu et al., 2025; Sun et al., 2026; Zhou et al., 2024; Feng et al., 2026) have advanced MLLM distillation by transferring complementary knowledge from strong teachers to compact students, including output distributions (Cai et al., 2025), multimodal representations (Xu et al., 2024; Jain et al., 2026), and token-level relations (Feng et al., 2025; Feng et al., 2026). These methods demonstrate that supervision beyond task labels can effectively preserve multimodal capability in smaller models. Nevertheless, they predominantly formulate distillation as matching static targets, such as final logits, selected hidden states, or token relations at isolated layers. This overlooks a key property of Transformer (Vaswani et al., 2017)-based MLLMs: cross-modal reasoning is not produced in one step, but evolves layer by layer through residual (He et al., 2016) updates as illustrated by Fig. 2.

Motivation. This layer-wise evolution naturally suggests a dynamical interpretation of MLLMs. Under the neural-ODE view (Gong et al., 2025), the layer index acts as discrete time and each residual block specifies a state-change direction, so an MLLM's forward pass is a continuous cross-modal evolution rather than a sequence of independent features. Accordingly, an MLLM does not process multimodal inputs as a sequence of independent layer-wise features; it continuously transforms an initial collection of visual and textual embeddings into progressively more task-relevant representations. In particular, deep-level multimodal semantic visual information is not simply preserved in a final hidden state. Through repeated cross-token interactions, language representations progressively select relevant visual evidence, suppress irrelevant content, and integrate the remaining evidence into semantic concepts across depth. The teacher's multimodal knowledge is therefore encoded both in its intermediate cross-modal grounding states and in the vector field that drives their evolution from layer to layer. Distilling only final logits or isolated representations offers sparse supervision for this inherently dynamic process. A more faithful student should approximate the teacher's cross-modal evolution itself, including both the intermediate grounding states and their layer-to-layer evolution.

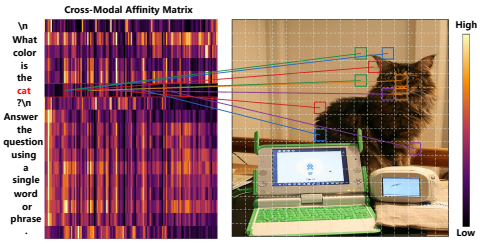


Figure 3: The student's cross-modal affinity matrix $A_{S,23}$ showing semantic grounding of "cat" to image regions.

Three fundamental challenges must be addressed to achieve effective MLLM distillation under practical constraints. 1) **How can knowledge be transferred across mismatched visual resolutions?** High-resolution teachers use substantially more visual tokens than constrained students, rendering token-wise relational supervision ill-defined without a compatible visual space. 2) **What knowledge should be aligned to preserve multimodal reasoning across heterogeneous models?** Direct hidden-state matching is sensitive to architectural and dimensional differences, while final logits only indirectly supervise visual grounding. The student instead requires a modality-aware signal linking text tokens to visual evidence. Fig. 3 illustrates this via the student's cross-modal affinity matrix. 3) **How can distillation capture the progressive nature of cross-modal reasoning?** Aligning affinities at isolated layers neglects how text-vision grounding is dynamically established and refined through successive residual updates. As empirically illustrated in Fig. 2, this progressive alignment enables the model to dynamically refine its visual focus: as layers deepen, the target token "cat" progressively allocates more attention to spatially correlated visual tokens.

Contributions. To this end, we propose CM-FDD, a cross-modal feature dynamics distillation framework for compact MLLMs in resource-constrained environments. CM-FDD introduces three key innovations to address the above challenges: 1) **Cross-Resolution Representation Projection (CRRP)** bridges the visual-resolution mismatch via deterministic bilinear projection. 2) **Cross-Modal Affinity Distillation (CMAD)** aligns teacher and student affinities at multiple depths, encouraging the student to reproduce the teacher's grounding states. 3) **Cross-Modal Affinity Evolution Distillation (CMAED)** aligns changes in affinity across layers, transferring the teacher's progressive grounding refinement. Grounding-state and grounding-evolution supervision impose complementary constraints.

Experiments systematically validate both the motivation and effectiveness of CM-FDD. As summarized in Fig. 1, CM-FDD consistently outperforms prior MLLM distillation methods across both model scales and achieves strong performance across ten multimodal benchmarks.

2 Related Work

Multimodal Large Language Models and Knowledge Distillation. Visual instruction tuning has substantially advanced MLLMs (Liu et al., 2023), from query-based or cross-attention fusion (Li et al., 2023a; Alayrac et al., 2022) to lightweight MLP projectors (Liu et al., 2024b; Cha et al., 2024), with high-resolution inputs further improving fine-grained perception (Lu et al., 2024; Kar et al., 2024). Nevertheless, MLLMs still struggle with fine-grained visual understanding (Tong et al., 2024; Qi et al., 2025), suffering attention misalignment (Jiang et al., 2025; Neo et al., 2025; Darcet et al., 2024) and visual information dilution across intermediate layers (Kaduri et al., 2025; Yoon et al., 2025; Chen et al., 2025), which motivates modeling how cross-modal grounding evolves through network depth. Knowledge distillation (KD) (Hinton et al., 2015) transfers knowledge from large teachers to compact students: early MLLM methods use teacher-generated data, logit alignment, and intermediate feature supervision (Zhou et al., 2024; Xu et al., 2024; Liu et al., 2024a; Shu et al., 2025), while later ones exploit visual relational knowledge (Cai et al., 2025), cross-modal alignment (Feng et al., 2025), multimodal representations (Xu et al., 2024; Jain et al., 2026), and visual attention or token interactions (Kim et al., 2026; Chen et al., 2026; Cao et al., 2025). Despite increasingly rich supervision (Sun et al., 2026), these methods largely remain snapshot-based, matching teacher knowledge at isolated stages.

Trajectory and Residual Dynamics Distillation Recent work views Transformers as discretized ODEs. Gong et al. (2025) matches intermediate prediction trajectories and their first-order derivatives across layers, while MTA (Chi et al., 2026) performs layer-adaptive alignment from word-level to phrase-level spans. A complementary line treats residual updates as the distillation signal: On et al. (2026) proposes to learn from student–teacher residual hidden states to avoid copying teacher errors, and DDL (Zhang et al., 2026) explicitly parameterizes depth-wise delta residual updates. TAKE (Vo et al., 2026) extends trajectory-aware selection to data distillation via influence scores during training. In contrast, we distill cross-modal feature dynamics to preserve both grounding states and their layer-to-layer evolution, enabling more complete transfer of the teacher's cross-modal grounding behavior.

3 Methodology

To transfer cross-modal reasoning from a large teacher to a compact student, we formulate CM-FDD around two complementary forms of depth-wise grounding knowledge: the *grounding state* at each matched depth and the *grounding evolution* between consecutive matched depths. Let $\mathcal{M}_T$ and $\mathcal{M}_S$ denote the teacher and student model, respectively. The teacher and student may differ in model depth, hidden dimension, and visual-token resolution.

Let $\{(s_k, t_k)\}_{k=1}^{K}$ denote K student–teacher decoder-layer pairs matched by relative depth and ordered from shallow to deep. At each pair, we represent text-to-vision grounding by cross-modal affinity state, $\mathbf{A}_{S,k}$ and $\mathbf{A}_{T,k}$. The corresponding finite-difference changes, $\Delta\mathbf{A}_{S,k}$ and $\Delta\mathbf{A}_{T,k}$, describe how grounding is revised between two consecutive matched checkpoints. CM-FDD therefore

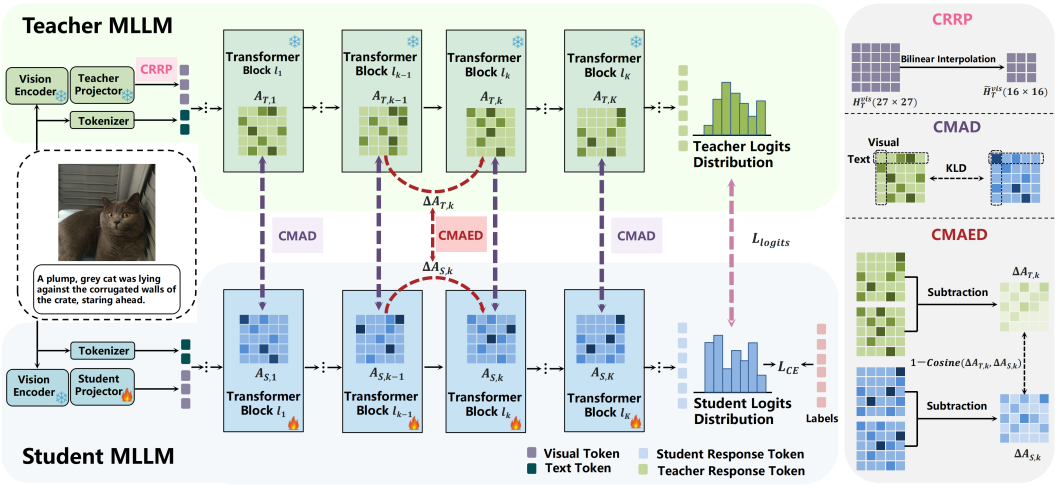


Figure 4: **Overall framework of CM-FDD.** We represent cross-modal grounding with text-to-vision affinity and distill it from two complementary perspectives. CMAD aligns grounding states at matched depths (Sec. 3.2), whereas CMAED aligns their cross-depth transitions to transfer progressive grounding refinement (Sec. 3.3). For teacher–student pairs with different visual-token grids, CRRP establishes a common visual support for affinity-based supervision (Sec. 3.1). CM-FDD thus transfers both where visual evidence is grounded and how that grounding evolves across depth.

transfers teacher knowledge through

$$\underbrace{\{\mathbf{A}_{T,k}\}_{k=1}^{K}}_{\text{grounding states}} + \underbrace{\{\Delta\mathbf{A}_{T,k}\}_{k=2}^{K}}_{\text{grounding evolution}} \longrightarrow \mathcal{M}_S. \tag{1}$$

As illustrated in Fig. 4, CMAD (Sec. 3.2) aligns the grounding states, while CMAED (Sec. 3.3) aligns their transitions. When the teacher and student use different visual grids, CRRP (Sec. 3.1) establishes a common visual support on which these affinity distributions can be compared. We first construct this model-comparable grounding representation, and then introduce grounding-state and grounding-evolution distillation in turn.

3.1 Cross-Modal Grounding Representation

Layer correspondence and hidden-state extraction. We select K matched pairs by relative depth, with denser coverage in early-to-middle layers. For the k-th matched pair (s_k, t_k), we extract text and visual hidden states $H_{S,k}^{text}, H_{S,k}^{vis}$ from the student and $H_{T,k}^{text}, H_{T,k}^{vis}$ from the teacher. Here, k indexes matched checkpoints rather than the absolute Transformer-block index of either model.

Cross-resolution compatibility via CRRP. The teacher and student encode the image into N_T^{vis} and N_S^{vis} visual tokens, corresponding to spatial grids $G_T^h \times G_T^w$ and $G_S^h \times G_S^w$. When $N_T^{vis} \neq N_S^{vis}$, their affinity distributions would otherwise be defined over different visual supports. As illustrated in Fig. 4, we therefore project the teacher visual states onto the student grid using a deterministic, parameter-free bilinear operator:

$$\widetilde{H}_{T,k}^{vis} = \mathcal{P}_{T\to S}(H_{T,k}^{vis}), \quad \text{where } \widetilde{H}_{T,k}^{vis}(u) = \sum_{p\in\mathcal{N}(u)} \alpha_{u,p} H_{T,k}^{vis}(p). \tag{2}$$

Here, u is a student-grid location, $\mathcal{N}(u)$ denotes its four nearest teacher-grid neighbors, and $\alpha_{u,p}$ are standard bilinear interpolation weights with $\alpha_{u,p} \geq 0$ and $\sum_p \alpha_{u,p} = 1$. We set $\widetilde{H}_{S,k}^{vis} = H_{S,k}^{vis}$.[1] CRRP thus serves as a compatibility mechanism that preserves spatial topology and local continuity while introducing no trainable parameters, which efficiently addresses the first question raised in the Introduction: **How can knowledge be transferred across mismatched visual resolutions?** Additional comparisons with alternative projection strategies are provided in Appendix E.

[1] When the two resolutions already match, $\mathcal{P}_{T\to S}$ degenerates into the identity mapping.

Text-to-vision affinity as grounding state. Since direct MSE matching is sensitive to heterogeneous feature scales and dimensions, we instead represent cross-modal interactions using text-to-vision affinity. For the k-th matched checkpoint, let $H^{\text{text}}_{m,k}$ and $\widetilde{H}^{\text{vis}}_{m,k}$ denote the selected text and projected visual hidden states of model $m \in \{S, T\}$. We first compute their pairwise cosine similarities:

$$C_{m,k}[b, n, q] = \cos\Big(H^{\text{text}}_{m,k}[b, n, :], \widetilde{H}^{\text{vis}}_{m,k}[b, q, :]\Big), \tag{3}$$

where n indexes a valid assistant-response position, q indexes a visual-token position, and b indexes the sample in the batch. We then normalize each selected text token's similarities over the visual-token dimension:

$$\mathbf{A}_{m,k}[b, n, q] = \text{Softmax}_q\left(\frac{C_{m,k}[b, n, q]}{\tau_{\text{aff}}}\right), \qquad \mathbf{A}_{m,k} \in \mathbb{R}^{B \times N^{\text{text}} \times N^{\text{vis}}_S}. \tag{4}$$

Here, N^{text} is the batch-padded maximum number of valid assistant-response positions. Thus, $\mathbf{A}_{m,k}[b, n, :]$ is a probability distribution over visual tokens that describes the visual regions most compatible with assistant-response position n. After CRRP, both $\mathbf{A}_{T,k}$ and $\mathbf{A}_{S,k}$ lie in a common distributional space with the same visual-token resolution. Each valid row specifies the relative affinity of one assistant-response position to the visual-token grid, thereby describing which visual evidence is emphasized at that depth. Because these affinities are computed within each model's own hidden space, their comparison does not require the teacher and student hidden dimensions to match. We regard $\mathbf{A}_{m,k}$ as the *grounding state* at the k-th matched depth.

3.2 Cross-Modal Affinity Distillation (CMAD): Grounding-State Alignment

CMAD asks where the student should ground visual evidence at each selected depth. Given the affinity states above, we align the teacher and student distributions with teacher-to-student KL divergence. For checkpoint k,

$$\mathcal{L}^k_{\text{state}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathcal{V}_b|}\sum_{n \in \mathcal{V}_b} D_{\text{KL}}(\mathbf{A}_{T,k}[b, n, :] \,\|\, \mathbf{A}_{S,k}[b, n, :]), \tag{5}$$

$\mathcal{V}_b$ is the set of valid assistant-response-token positions for sample b. Padding positions are excluded by the text mask, so only $n \in \mathcal{V}_b$ contribute to the loss, and the multi-depth grounding-state objective is

$$\mathcal{L}_{\text{state}} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}^k_{\text{state}}. \tag{6}$$

KL divergence emphasizes visual locations assigned high relevance by the teacher while retaining the uncertainty encoded by the full affinity distribution. Applying this supervision at multiple matched depths provides a sequence of grounding-state constraints rather than relying on an isolated final-layer snapshot, as illustrated by the depth-wise teacher–student state alignment in Fig. 4.

Importantly, CMAD provides explicit supervision over the grounding state at each selected depth, encouraging the student to recover the teacher's depth-wise text-to-vision grounding patterns. This directly addresses the second question posed in the Introduction: **What knowledge should be aligned to preserve multimodal reasoning across heterogeneous models?** Beyond static states, grounding is progressively refined across depth, motivating CMAED below.

3.3 Cross-Modal Affinity Evolution Distillation (CMAED)

CMAED asks how grounding should be revised from one selected depth to the next. For consecutive matched checkpoints, we define the teacher and student finite-difference grounding changes as

$$\Delta\mathbf{A}_{T,k} = \mathbf{A}_{T,k} - \mathbf{A}_{T,k-1} \qquad \Delta\mathbf{A}_{S,k} = \mathbf{A}_{S,k} - \mathbf{A}_{S,k-1}, \qquad k = 2, \ldots, K. \tag{7}$$

The change $\Delta\mathbf{A}_{m,k}$ records how visual evidence is strengthened, suppressed, or redistributed as model m moves from one selected depth to the next. Residual networks admit a depth-wise dynamical interpretation in which successive blocks implement discrete representation updates; under this view, the finite difference above serves as a discrete proxy for the local *direction* of grounding evolution rather than an exact continuous-time derivative. A fuller neural-ODE interpretation is deferred to Appendix A.

Because the teacher and student may differ in capacity and in the depth intervals between matched checkpoints, we align the direction of affinity changes rather than forcing their magnitudes to be identical. For each transition k, let $\mathcal{B}_k$ denote the samples containing at least one valid assistant-response token. After applying the valid sample/text mask, we treat each valid assistant-response token position $n \in \mathcal{V}_b$ separately, using its per-row affinity change $\Delta\mathbf{A}_{m,k,n,:}$ and its norm $a_{m,k,n} = \|\Delta\mathbf{A}_{m,k,n,:}\|_2$. The evolution-direction loss is then computed row by row with a stabilized cosine distance

$$\ell_{\text{evo}}^{k,n} = \begin{cases} 1 - \cos(\Delta\mathbf{A}_{T,k,n,:}, \Delta\mathbf{A}_{S,k,n,:}), & a_{T,k,n} > \epsilon \wedge a_{S,k,n} > \epsilon, \\ 1, & (a_{T,k,n} > \epsilon) \oplus (a_{S,k,n} > \epsilon), \\ 0, & \text{otherwise}, \end{cases}$$

where $\oplus$ denotes exclusive OR and the cosine denominator is guarded by $\max(a_{T,k,n}a_{S,k,n}, \epsilon)$. The final loss averages these per-row terms over all valid rows and consecutive layer transitions:

$$\mathcal{L}_{\text{evo}} = \frac{1}{K-1}\sum_{k=2}^{K} \frac{1}{\sum_{b\in\mathcal{B}_k}|\mathcal{V}_b|} \sum_{b\in\mathcal{B}_k}\sum_{n\in\mathcal{V}_b} \ell_{\text{evo}}^{k,n}, \qquad \epsilon = 10^{-8}. \tag{8}$$

This per-row treatment detects degenerate transitions locally, rather than absorbing them into a sample-level cosine, while still aligning the direction rather than the magnitude of affinity changes. Invalid rows and samples contribute no loss, and weighting is per valid row rather than per sample, so every valid response token counts equally and a longer response weighs proportionally more. It encourages the student to reproduce the teacher's redistribution of visual affinity for each answer token(see Appendix D.2 for a sample-level variant loss).

By aligning these per-row evolution directions, CMAED provides explicit supervision over the teacher's depth-wise grounding refinement, encouraging the student to strengthen, suppress, and redistribute visual evidence in a teacher-consistent manner across depths, as illustrated in Fig. 4. This directly addresses the third question posed in the Introduction: **How can distillation capture the progressive nature of cross-modal reasoning?** By jointly imposing the two complementary constraints, we obtain the grounding-dynamics objective

$$\mathcal{L}_{\text{dyn}} = \mathcal{L}_{\text{state}} + \beta\mathcal{L}_{\text{evo}}, \tag{9}$$

where β balances state and evolution supervision. The state term constrains where the teacher grounds visual evidence, whereas the evolution term constrains how that grounding is progressively refined across depth.

3.4 Three-Stage Distillation Training Pipeline

We adopt a three-stage training procedure designed to progressively establish visual-interface compatibility, adapt the student to multimodal instruction following, and finally transfer the teacher's grounding dynamics. Given an image–text instruction pair (I_b, X_b) and target response $Y_b = (y_{b,1}, \ldots, y_{b,L_b})$, we define the standard autoregressive cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the logit distillation loss $\mathcal{L}_{\text{logit}}$ as

$$\mathcal{L}_{\text{CE}} = -\frac{1}{\sum_{b\in\mathcal{B}}|\mathcal{V}_b|}\sum_{b\in\mathcal{B}}\sum_{i\in\mathcal{V}_b} \log p_S(y_{b,i} \mid I_b, X_b, y_{b,<i})\,. \tag{10}$$

$$\mathcal{L}_{\text{logit}} = \frac{1}{\sum_{b\in\mathcal{B}}|\mathcal{V}_b|}\sum_{b\in\mathcal{B}}\sum_{i\in\mathcal{V}_b} D_{\text{KL}}\left(\text{Softmax}\left(\frac{z_{T,b}^i}{\tau_{\text{logit}}}\right)\middle\|\text{Softmax}\left(\frac{z_{S,b}^i}{\tau_{\text{logit}}}\right)\right)\tau_{\text{logit}}^2, \tag{11}$$

where $z_{T,b}^i$ and $z_{S,b}^i$ are the teacher and student logits at position i of sample b, and τ_{logit} is the logit temperature. Both losses are computed only over the valid assistant-response-token positions $\mathcal{V}_b$.

Stage 1: Distilled Pre-Training (DPT). DPT is an interface-alignment stage. We freeze the student LLM, teacher, and vision encoders and optimize only the student projector. Because the Transformer itself is frozen, this stage uses grounding-state supervision to encourage the projected visual input to induce teacher-compatible grounding inside the student LLM. Using a sparse subset of matched checkpoints $\mathcal{J}_{\text{DPT}}$ spanning early, middle, and late stages, we optimize

$$\mathcal{L}_{\text{DPT}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{dpt}}\left(\frac{1}{|\mathcal{J}_{\text{DPT}}|}\sum_{k\in\mathcal{J}_{\text{DPT}}}\mathcal{L}_{\text{state}}^k\right). \tag{12}$$

Table 1: Comparison with compact MLLMs at the 1B and 2B scales. The upper and lower sections report the 1B and 2B models, respectively. "–" denotes an unavailable result. Avg_7: Average of seven benchmarks (excluding VQAv2, POPE, MMMU) for direct comparison with existing MLLM distillation methods. Avg_{10}: Average across all benchmarks for comprehensive comparison. †: reproduced results using the official code. Best and second-best results are shown in **bold** and underlined, respectively.

Method	LLM	Res.	Image Question Answering					Benchmarks					Avg_7	Avg_{10}
			VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU		
SPHINX-Tiny	TinyLlama-1.1B	–	74.7	58.0	**49.2**	21.5	57.8	63.1	52.3	56.6	82.2	–	51.2	–
EM-KD	Qwen2.5-0.5B	224	74.8	58.3	28.9	59.1	49.2	61.5	58.9	54.2	86.1	**33.6**	52.9	56.5
TinyLLaVA	TinyLlama-1.1B	384	75.5	58.6	–	**64.0**	49.6	62.8	–	–	86.3	28.3	–	–
LLaVADI	MLLaMA-1.4B	336	–	55.4	–	56.0	45.3	58.9	55.0	–	84.7	–	–	–
LLaVA-MOD	Qwen2-0.5B	336	–	56.6	35.1	61.8	57.1	67.0	58.7	54.1	–	–	55.7	–
COMPODISTILL	Qwen1.5-0.5B	384	–	60.1	–	–	50.3	59.5	–	–	–	–	–	–
Align-TI†	Qwen2-0.5B	336	**78.4**	**60.4**	32.8	60.7	**59.9**	68.9	63.2	**66.7**	86.8	29.1	58.9	60.7
LLaVA-KD	Qwen2.5-0.5B	384	77.7	59.8	41.5	60.6	52.0	64.7	61.3	57.0	86.4	28.3	56.7	58.9
CM-FDD (Ours)	Qwen2.5-0.5B	224	71.1	56.2	42.0	60.1	45.0	72.2	66.4	62.9	85.7	32.2	57.8	59.4
CM-FDD (Ours)	Qwen2.5-0.5B	384	75.4	59.0	42.6	59.0	51.0	**75.2**	**68.7**	65.4	**87.1**	31.8	**60.1**	**61.5**
LLaVA-1.5	Vicuna-7B	336	78.5	62.0	50.0	66.8	58.2	75.5	64.3	58.3	85.9	34.4	62.2	63.4
TinyLLaVA	Qwen2.5-1.5B	384	78.8	62.0	43.2	72.0	57.4	72.5	68.6	63.0	85.5	**37.0**	62.7	64.0
LLaVADI	MLLaMA-2.7B	336	–	61.4	–	64.1	50.7	68.8	62.5	–	86.7	–	–	–
LLaVA-MOD	Qwen2-1.5B	336	–	58.8	40.4	69.2	59.9	69.2	68.9	64.4	–	–	61.6	–
COMPODISTILL	Qwen1.5-1.8B	384	78.8	62.2	42.5	–	56.4	70.1	–	–	–	–	–	–
MOVE-KD	MLLaMA-1.4B	336	71.5	57.7	28.6	56.0	42.6	57.3	47.0	–	84.3	–	–	–
Beta-KD†	MLLaMA-1.4B	336	76.9	61.2	44.5	62.9	54.2	67.5	60.2	49.8	86.0	30.7	57.2	59.4
Align-TI†	Qwen2-1.5B	336	**80.9**	**62.9**	36.9	71.4	**65.1**	75.6	71.8	66.7	86.1	29.1	64.3	64.7
LLaVA-KD	Qwen2.5-1.5B	384	80.3	62.5	46.0	71.6	59.7	70.0	71.0	66.6	86.7	35.8	63.9	65.0
CM-FDD (Ours)	Qwen2.5-1.5B	224	76.6	61.0	**51.2**	73.2	54.1	82.3	**77.4**	74.5	86.6	35.8	67.7	67.3
CM-FDD (Ours)	Qwen2.5-1.5B	384	79.4	62.3	48.5	**73.4**	60.3	**86.7**	77.3	**76.1**	**88.3**	36.3	**69.2**	**68.9**

This encourages the projector to induce teacher-compatible text-to-vision grounding throughout the frozen LLM, from early multimodal fusion to late semantic refinement.

Stage 2: Supervised Fine-Tuning (SFT). Starting from the DPT checkpoint, we fine-tune the student LLM and projector on multimodal instruction data while keeping the vision encoder frozen. No teacher constraint is imposed in this stage, which adapts the student to multimodal instruction following:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}}. \tag{13}$$

Stage 3: Cascaded Distillation Fine-Tuning (CDFT). Once the student Transformer is trainable, we apply the complete state-and-evolution supervision together with response-level distillation. The CDFT objective is

$$\mathcal{L}_{\text{CDFT}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{dyn}}(r)\mathcal{L}_{\text{dyn}} + \lambda_{\text{logit}}(r)\mathcal{L}_{\text{logit}}, \tag{14}$$

where $r \in [0, 1]$ is normalized CDFT training progress. Here, $\mathcal{L}_{\text{state}}$ transfers *where* to ground, $\mathcal{L}_{\text{evo}}$ transfers *how* grounding is refined, and $\mathcal{L}_{\text{logit}}$ transfers the final response distribution, complementing the intermediate grounding-process supervision. We use the linear schedule

$$\lambda_{\text{dyn}}(r) = \lambda_{\text{dyn}}^{\max}\left[1 - 0.5\min\left(\frac{r}{0.4}, 1\right)\right], \qquad \lambda_{\text{logit}}(r) = \lambda_{\text{logit}}^{\max}\left[0.1 + 0.9\min\left(\frac{r}{0.4}, 1\right)\right]. \tag{15}$$

During the first 40% of CDFT, grounding-dynamics supervision gradually decreases from 1.0 to 0.5 of its maximum weight, while logit supervision increases from 0.1 to 1.0; both remain constant thereafter. This schedule places relatively stronger emphasis on intermediate grounding transfer early in CDFT and increasingly consolidates output-distribution matching later.

4 Experiments

4.1 Experimental Settings

Implementation Details. For both the teacher and student, we employ the SigLIP2 vision encoder (Tschannen et al., 2025) and the Qwen2.5 LLM family. We train student LLMs at two different scales, namely Qwen2.5-0.5B and Qwen2.5-1.5B, under two different resolutions (224×224 and 384×384). The teacher model is Qwen2.5-3B, solely under 384×384 resolution. In the DPT stage,

we used LLaVA1.5-558k (Liu et al., 2024b) as the training set, while in the SFT and CDFT stages, we used the LLaVA-mix-665k (Liu et al., 2024b) dataset for fine-tuning. The loss parameters for the three training stages, λ_{dpt}, $\lambda_{\text{dyn}}^{\max}$, $\lambda_{\text{logit}}^{\max}$, were set to 100, 0.3, 0.65, respectively. Detailed training parameter settings can be found in Appendix B.

Evaluation Benchmarks. Following the setup of LLaVA-KD, for general VQA tasks, we use VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), and ScienceQA (Lu et al., 2022) to test the model's ability to answer questions when image and text information are interrelated; we use TextVQA (Singh et al., 2019) to test the model's ability to recognize textual information within images; and we use MME (Fu et al., 2026), MMBench (Liu et al., 2024c), MMBench_CN (Liu et al., 2024c), POPE (Li et al., 2023b), and MMMU (Yue et al., 2024) to test the model's comprehensive capabilities.

4.2 Quantitative Results on Ten Benchmarks

As shown in Table 1, CM-FDD achieves state-of-the-art performance across 1B and 2B scales, outperforming LLaVA-KD (Cai et al., 2025), Align-TI (Chen et al., 2026), Align-KD (Feng et al., 2025), COMPODISTILL (Kim et al., 2026), TinyLLaVA (Zhou et al., 2024), LLaVA-MOD (Shu et al., 2025), LLaVADI (Xu et al., 2024), Beta-KD (Sun et al., 2026), and SPHINX-Tiny (Liu et al., 2024a). With the same data, CM-FDD surpasses LLaVA-KD by 2.6% and 3.9% on Avg_{10} at 1B and 2B scales, respectively. This supports three key findings: (1) The CRRP module effectively handles cross-resolution distillation between heterogeneous models, enabling the 224×224 student to maintain strong performance with a 384×384 teacher. (2) The method scales robustly across model sizes. (3) With only 1.2M data (LLaVA1.5-558k + LLaVA-mix-665k), CM-FDD outperforms models trained on much larger datasets (e.g., LLaVA-MOD with 5M, Align-TI with 3.6M), highlighting its data efficiency and promising direction for efficient MLLM distillation.

Table 2: Ablation for CRRP.

Downsample Methods	Avg_{10}
Lightweight Refinement Module	57.0
Average Pooling	58.2
Hungarian Matching	58.2
Bilinear Interpolation	**59.4**

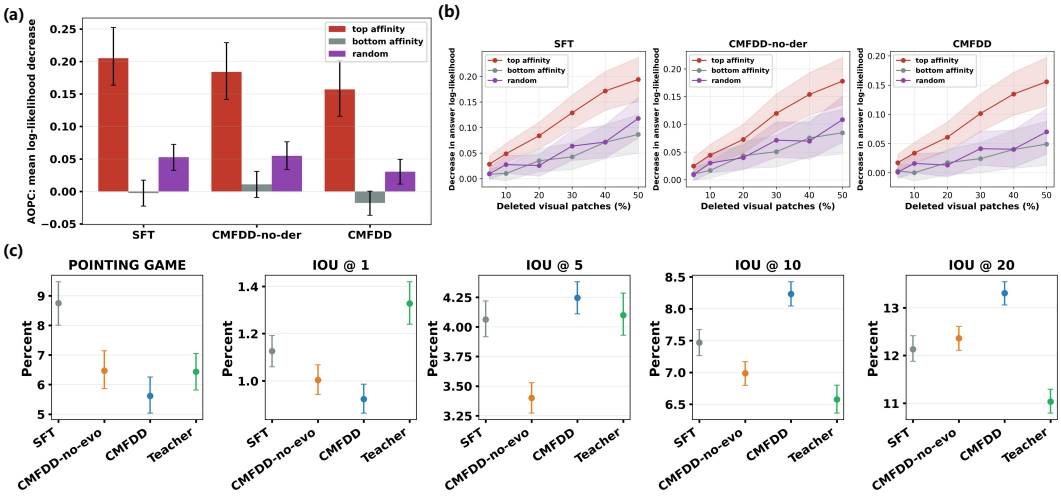


Figure 5: (a) Causal faithfulness (bars are comparable within a model, not across models). (b) Causal visual-evidence deletion on VQAv2. (c) RefCOCO testA localization.

4.3 Grounding and Causal Faithfulness Evaluation

We evaluate cross-modal affinities by causal deletion (VQAv2) and localization (RefCOCO (Yu et al., 2016)). In Fig. 5(a,b), blurring patches ranked by affinity yields the area over the deletion curve (AOPC). Within every checkpoint the top-affinity patches dominate the controls, so the map locates evidence the model uses; since the absolute drop also scales with each model's perturbation

sensitivity, we read AOPC as a within-model contrast and verify in (b) that the ordering persists under the teacher's ranking. In (c), CM-FDD attains higher IoU over the top 5–20% of affinity mass, covering the referred region rather than a single cell, and accordingly trails on Pointing Game and IoU@1. These contrasts, which are weaker for the w/o $\mathcal{L}_{\text{evo}}$ variant, support modeling grounding dynamics across depth.

4.4 Ablation Studies

Effect of each module. We conduct ablation studies to verify the effectiveness of each module. Notably, for training efficiency, all ablation experiments are conducted using the Qwen2.5-0.5B student model with a 224×224 input resolution. Table 2 presents the performance of CRRP compared with other downsampling methods, while Table 3 illustrates the contributions of CMAD and CMAED. From Table 2, it can be observed that the model adopting the Bilinear Interpolation strategy achieves the best overall performance, as it preserves spatial consistency among visual tokens and introduces no trainable modules. A detailed analysis of the other three downsampling methods is provided in Appendix E. Table 3 demonstrates that merely distilling cross-layer affinities is insufficient; only through the combined effect of CMAD and CMAED can the optimal knowledge distillation be achieved for the student.

Table 3: Ablation for CMAD and CMAED.

CMAD module	CMAED module	Avg_{10}
×	×	58.0
✓	×	59.1
×	✓	57.9
✓	✓	**59.4**

Effect of the Number of Distilled Layers. Fig. 6 varies the number of selected student–teacher layer pairs K from 3 to 14 at both student scales. Performance remains highly stable across this range: the Qwen2.5-0.5B student fluctuates within only 0.16 points (59.22–59.38) and the Qwen2.5-1.5B student within 0.23 points (67.19–67.42), with no monotonic trend. This insensitivity indicates that CM-FDD does not rely on a finely tuned sampling density: a few well-spread layer pairs already suffice to characterize the cross-modal grounding trajectory, while denser sampling only adds redundant but harmless supervision. The flat profiles at both scales confirm that the distilled dynamics signal is stable and that the method is robust to the choice of K, reducing the burden of hyperparameter tuning in practice (see Appendix B for the specific layer choices).

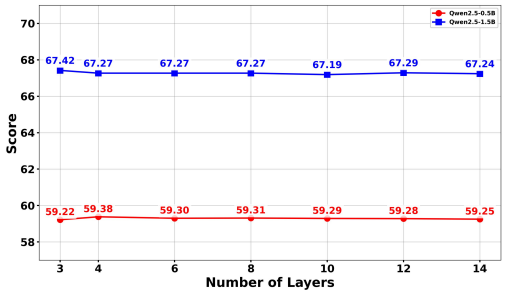


Figure 6: Effect of the number of selected student layers K for cross-modal distillation.

Effect of Teacher and Student Sizes. Table 4 examines CM-FDD under two student capacities, Qwen2.5-0.5B and Qwen2.5-1.5B, distilled from Qwen2.5-7B or Qwen2.5-3B teachers at a fixed resolution of 224. This comparison isolates the effect of teacher capacity and the teacher–student capacity gap on cross-modal dynamics transfer. The results reveal that the closer 3B teacher yields a more favorable distillation target for the compact student, as it provides more useful supervision than the 7B teacher (see Appendix C for further discussion).

Table 4: Effect of teacher and student model sizes.

Teacher	Student	Res.	Avg_7	Avg_{10}
Qwen2.5-7B	Qwen2.5-0.5B	224	57.5	59.3
Qwen2.5-3B	Qwen2.5-0.5B	224	**57.8**	**59.4**
Qwen2.5-7B	Qwen2.5-1.5B	224	66.8	66.9
Qwen2.5-3B	Qwen2.5-1.5B	224	**67.7**	**67.3**

5 Conclusion

We presented CM-FDD, which distills cross-modal reasoning dynamics, including grounding states and their layer-wise evolution, from large MLLMs to compact students. Experiments across ten benchmarks confirm state-of-the-art results at 1B and 2B scales. These results show that aligning the evolution of cross-modal grounding provides more effective supervision than distilling only final outputs or isolated representations.

6 AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods, providing feedback on experiments, translating text, and interpreting results. We did not use them to generate synthetic datasets, write proofs, format references, create or modify scientific figures or images, create artifacts, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proofs, propose or refine hypotheses, clean or reformat datasets, or support qualitative and thematic data analysis. We also used such tools for brainstorming, identifying research gaps, and implementing methods. All AI-assisted work was reviewed. LLM-generated code was verified and tested for correctness by two authors, and LLM-generated research ideas were checked for potential plagiarism through a manual literature survey. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 239–249, 2025.

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 19846–19856. IEEE, 2025.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 13817–13827, 2024.

Lin Chen, Xiaoke Zhao, Kun Ding, Weiwei Feng, Changtao Miao, Zili Wang, Wenxuan Guo, Ying Wang, Kaiyuan Zheng, Bo Zhang, et al. Beyond next-token alignment: Distilling multimodal large language models via token interactions. In *Forty-third International Conference on Machine Learning, ICML*, 2026.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *Forty-Second International Conference on Machine Learning, ICML*, 2025.

Pham Khanh Chi, Quoc Phong Dao, Thuật Nguy˜ên, Linh Ngo Van, Trung Le, and Thanh Hong Nguyen. Mta: Multi-granular trajectory alignment for large language model distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32669–32684, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations, ICLR*, volume 2024, pp. 2632–2652, 2024.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 4178–4188, 2025.

Ze Feng, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. Em-kd: distilling efficient multimodal large language model with unbalanced vision tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21111–21119, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Guoqiang Gong, Jiaxing Wang, Jin Xu, Deping Xiang, Zicheng Zhang, Leqi Shen, Yifeng Zhang, JunhuaShu JunhuaShu, ZhaolongXing ZhaolongXing, Zhen Chen, et al. Beyond logits: Aligning feature dynamics for effective knowledge distillation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23067–23077, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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, CVPR*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 3608–3617, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Drew A Hudson and Christopher D Manning. 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, CVPR*, pp. 6700–6709, 2019.

Jitesh Jain, Zhengyuan Yang, Humphrey Shi, Jianfeng Gao, and Jianwei Yang. Elevating visual perception in multimodal llms with visual embedding distillation. *Advances in Neural Information Processing Systems*, 38:82020–82051, 2026.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 25004–25014, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image$f$ a deep-dive into the vision of vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 14549–14558, 2025.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision, ECCV*, 2024.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. Compodistill: Attention distillation for compositional reasoning in multimodal llms. *The Fourteenth International Conference on Learning Representations, ICLR*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Fortieth International Conference on Machine Learning, ICML*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pp. 292–305, 2023b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 26689–26699, 2024.

Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *Forty-First International Conference on Machine Learning, ICML*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, CVPR*, pp. 26296–26306, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision, ECCV*, pp. 216–233. Springer, 2024c.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR*, volume 2025, pp. 57172–57189, 2025.

Thinh On, Hengzhi Pei, Leonard Lausen, and George Karypis. Knowledge distillation for large language models through residual learning. In *The Fourteenth International Conference on Learning Representations, ICLR*, 2026.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Weilong Dai, Zhelun Yu, Wanggui He, et al. Llava-mod: Making llava tiny via moe-knowledge distillation. In *The Thirteenth International Conference on Learning Representations, ICLR*, volume 2025, pp. 9386–9404, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 8317–8326, 2019.

Jingchen Sun, Shaobo Han, Deep Patel, Wataru Kohno, Can Jin, and Changyou Chen. Uncertainty-aware knowledge distillation for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 5585–5595, 2026.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 9568–9578, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Tri-Nhan Vo, Dang Nguyen, and Sunil Gupta. Take: Trajectory-aware knowledge estimation for text dataset distillation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 177–194. Springer, 2026.

Haojun Xia, Zhen Zheng, Yuchao Li, Donglin Zhuang, Zhongzhu Zhou, Xiafei Qiu, Yong Li, Wei Lin, and Shuaiwen Leon Song. Flash-llm: Enabling cost-effective and highly-efficient large generative model inference with unstructured sparsity. *arXiv preprint arXiv:2309.10285*, 2023.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision, ECCV*, pp. 69–85. Springer, 2016.

Zhongzhi Yu, Zheng Wang, Yuhan Li, Ruijie Gao, Xiaoya Zhou, Sreenidhi Reddy Bommu, Yang Zhao, and Yingyan Lin. Edge-llm: Enabling efficient large language model adaptation on edge devices via unified compression and adaptive layer voting. In *Proceedings of the 61st ACM/IEEE Design Automation Conference, DAC*, pp. 1–6, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 9556–9567, 2024.

Mingjin Zhang, Xiaoming Shen, Jiannong Cao, Zeyang Cui, and Shan Jiang. Edgeshard: Efficient llm inference via collaborative edge computing. *IEEE Internet of Things Journal*, 12(10):13119–13131, 2024.

Yifan Zhang, Yifeng Liu, Mengdi Wang, and Quanquan Gu. Deep delta learning. *arXiv preprint arXiv:2601.00417*, 2026.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

Appendix

A Mathematical Interpretation of CMAED 14
B Additional Implementation Details 16
C Teacher Models: Training, Evaluation, and Distillation Behavior 18
D Supplementary Ablation Study 18
D.1 All-Token Affinity Supervision 18
D.2 Sample-Level Affinity-Derivative Cosine Loss 19
D.3 Ablation study on Training Recipes 19
D.4 Effect of the LLM Backbone 20
D.5 Statistical Robustness Across Training Seeds 20
E Additional Analysis of Visual Token Downsampling Methods 21
F Diagnostic Probes of Grounding Transfer 22
F.1 Depth-Wise Affinity Dynamics Converge on the Teacher 22
F.2 Causal Necessity and Redundancy of the Transferred Grounding 24
G Inference Cost on Consumer Hardware 26
H Qualitative Case Studies 27

A Mathematical Interpretation of CMAED

This appendix provides a mathematical interpretation of the Cross-Modal Affinity Evolution Distillation (CMAED) module from the viewpoint of ordinary differential equations. The key observation is that the hidden states produced by successive Transformer blocks should be regarded as points sampled from a continuous representation flow, rather than as independent layer-wise features (Gong et al., 2025).

A.1 Transformer Blocks as a Discretized Dynamical System

Let $\mathbf{H}_m^\ell$ denote the hidden states at layer ℓ of model $m \in \{S, T\}$, where the hidden states contain both text and visual tokens. A residual Transformer block can be abstractly written as

$$\mathbf{H}_m^{\ell+1} = \mathbf{H}_m^\ell + \Delta t_\ell \mathbf{F}_m^\ell \left(\mathbf{H}_m^\ell\right), \tag{16}$$

where $\mathbf{F}_m^\ell$ represents the transformation induced by the self-attention and feed-forward sublayers, and Δt_ℓ is the effective depth interval between two consecutive states. Eq. 16 is the forward-Euler discretization of the neural ordinary differential equation

$$\frac{d\mathbf{H}_m(t)}{dt} = \mathbf{F}_m\left(\mathbf{H}_m(t), t\right). \tag{17}$$

Under this interpretation, the forward propagation of an MLLM is a dynamical system evolving along network depth. Its final hidden state is determined not only by the initial visual-language representation, but also by the sequence of transformations applied throughout the intermediate layers. Consequently, a student that matches only the final output may arrive at a similar endpoint through a substantially different internal trajectory.

A.2 Cross-Modal Affinity as a Dynamical State

At a selected layer, we separate the hidden states into text and visual parts:

$$\mathbf{H}_m(t) = \left[\mathbf{H}_m^{text}(t); \mathbf{H}_m^{vis}(t)\right]. \tag{18}$$

When the teacher and student have different visual resolutions, the teacher visual states are first projected onto the student visual grid. We then define the cross-modal affinity state as

$$\mathbf{A}_m(t) = \Phi\left(\mathbf{H}_m^{text}(t), \widetilde{\mathbf{H}}_m^{vis}(t)\right), \tag{19}$$

where Φ applies cosine similarity followed by a softmax over visual tokens. For a text token n, the corresponding affinity distribution is

$$\mathbf{A}_{m,n,q}(t) = \frac{\exp\left(\cos\left(\mathbf{h}_{m,n}^{text}(t), \widetilde{\mathbf{h}}_{m,q}^{vis}(t)\right)/\tau_{\text{aff}}\right)}{\sum_{q'=1}^{N_S^{vis}} \exp\left(\cos\left(\mathbf{h}_{m,n}^{text}(t), \widetilde{\mathbf{h}}_{m,q'}^{vis}(t)\right)/\tau_{\text{aff}}\right)}. \tag{20}$$

Thus, $\mathbf{A}_m(t)$ is not an additional hidden state explicitly propagated by the Transformer. Rather, it is an observable relational state induced by the evolving text and visual representations. It describes how the semantic content of each text token is distributed over the available visual positions at depth t. As the hidden states evolve according to Eq. 17, the affinity state also evolves:

$$\frac{d\mathbf{A}_m(t)}{dt} = \frac{\partial\Phi}{\partial\mathbf{H}_m^{text}}\frac{d\mathbf{H}_m^{text}(t)}{dt} + \frac{\partial\Phi}{\partial\widetilde{\mathbf{H}}_m^{vis}}\frac{d\widetilde{\mathbf{H}}_m^{vis}(t)}{dt}. \tag{21}$$

Substituting Eq. 17 into Eq. 21 gives

$$\frac{d\mathbf{A}_m(t)}{dt} = \frac{\partial\Phi}{\partial\mathbf{H}_m^{text}}\mathbf{F}_m^{text} + \frac{\partial\Phi}{\partial\widetilde{\mathbf{H}}_m^{vis}}\widetilde{\mathbf{F}}_m^{vis}. \tag{22}$$

This equation shows that the evolution of cross-modal affinity is jointly determined by the language-side transformation, the visual-side transformation, and their interaction through the affinity function. Therefore, affinity evolution provides a compact description of how multimodal representations progressively establish, modify, and refine associations between language and visual content.

A.3 Discrete Affinity Trajectory

In practice, we observe the continuous process only at K selected layer pairs. Let the selected student and teacher layers be uniformly indexed as

$$\mathcal{L}_S = \{S_1, S_2, \ldots, S_K\}, \qquad \mathcal{L}_T = \{T_1, T_2, \ldots, T_K\}, \tag{23}$$

where the shared indices indicate correspondence in the ordered trajectories, rather than identical physical layer numbers. The sampled affinity trajectories are

$$\mathcal{A}_S = \{\mathbf{A}_{S,1}, \mathbf{A}_{S,2}, \ldots, \mathbf{A}_{S,K}\}, \qquad \mathcal{A}_T = \{\mathbf{A}_{T,1}, \mathbf{A}_{T,2}, \ldots, \mathbf{A}_{T,K}\}. \tag{24}$$

The state loss aligns the teacher and student at every selected depth:

$$\mathcal{L}_{\text{state}} = \frac{1}{KB}\sum_{k=1}^{K}\sum_{b=1}^{B}\frac{1}{|\mathcal{V}_b|}\sum_{n\in\mathcal{V}_b} D_{\text{KL}}\left(\mathbf{A}_{T,k}[b,n,:] \,\|\, \mathbf{A}_{S,k}[b,n,:]\right), \tag{25}$$

where the KL divergence is computed over the visual-token dimension and averaged over valid text-token positions. This supervision constrains the student to pass through teacher-compatible intermediate cross-modal states instead of matching only the final state.

A.4 Affinity Derivatives and Evolution Directions

For two consecutive selected states, the affinity derivative is approximated by the finite difference

$$\dot{\mathbf{A}}_{m,k} \approx \frac{\mathbf{A}_{m,k} - \mathbf{A}_{m,k-1}}{\Delta t_k}, \qquad m \in \{S, T\}, \tag{26}$$

where Δt_k is the effective depth interval between the two selected layers. In the implementation, we use the unnormalized difference

$$\Delta\mathbf{A}_{m,k} = \mathbf{A}_{m,k} - \mathbf{A}_{m,k-1}. \tag{27}$$

The omission of Δt_k does not change the direction of the derivative when the layer interval is positive:

$$\cos\left(\frac{\Delta \mathbf{A}_{T,k}}{\Delta t_{T,k}}, \frac{\Delta \mathbf{A}_{S,k}}{\Delta t_{S,k}}\right) = \cos\left(\Delta \mathbf{A}_{T,k}, \Delta \mathbf{A}_{S,k}\right). \tag{28}$$

Because CMAED matches derivative directions through cosine similarity, the common positive scale factor cancels. The unnormalized difference also provides a consistent discrete measure across corresponding trajectory intervals and avoids introducing an additional depth-normalization factor.

We align the layer-to-layer evolution directions with a stabilized row-wise cosine loss. The evolution loss averages these per-row terms over all valid response-token rows and consecutive layer transitions:

$$\mathcal{L}_{\text{evo}} = \frac{1}{K-1}\sum_{k=2}^{K} \frac{1}{\sum_{b\in\mathcal{B}_k}|\mathcal{V}_b|} \sum_{b\in\mathcal{B}_k}\sum_{n\in\mathcal{V}_b} \ell_{\text{evo}}^{k,n}, \qquad \epsilon = 10^{-8}, \tag{29}$$

where $\mathcal{B}_k$ collects the samples containing at least one valid assistant-response token and $\mathcal{V}_b$ is the set of such positions in sample b; the main text abbreviates the term as $\ell_{\text{evo}}^{k,n}$ with the sample index suppressed. Rows excluded by the text mask do not enter the average at all, whereas stationary rows, whose two affinity changes both fall below ϵ, enter with value zero. Weighting is consequently per valid row rather than per sample: every valid response token contributes with equal weight, and a longer response carries proportionally more weight in the objective. Combining state and derivative supervision gives the CMAED objective:

$$\mathcal{L}_{\text{dyn}} = \mathcal{L}_{\text{state}} + \beta \mathcal{L}_{\text{evo}}, \tag{30}$$

where β controls the relative contribution of evolution-direction alignment. The state term constrains the locations of the sampled affinity states, whereas the derivative term constrains the directions in which those states move between layers.

A.5 Why Final-Layer Matching Is Insufficient

Matching only the final affinity matrix imposes the terminal constraint

$$\mathbf{A}_{S,k} \approx \mathbf{A}_{T,k}, \tag{31}$$

but places no constraint on $\mathbf{A}_S(t)$ for $t < K$. Many distinct trajectories can satisfy Eq. 31; consequently, a student may reach a similar final grounding distribution through an evolution that is inconsistent with the teacher. Enforcing only this endpoint reduces supervision to rote imitation of the teacher's terminal features. It forces the student to reproduce the teacher's final grounding distribution without the cross-modal computation that produces it, rather than developing its own reasoning trajectory. This compromises the transfer we ultimately care about.

By contrast, CMAED approximates both the state and vector field of the teacher:

$$\mathbf{A}_S(t_k) \approx \mathbf{A}_T(t_k), \qquad \frac{d\mathbf{A}_S}{dt}(t_k) \approx \frac{d\mathbf{A}_T}{dt}(t_k). \tag{32}$$

The first condition transfers the teacher's intermediate cross-modal grounding states, while the second transfers the layer-wise changes that establish and refine those states. Hence, the student is encouraged to reproduce not merely where the teacher's cross-modal reasoning ends, but also the process by which it develops across network depth.

B Additional Implementation Details

CM-FDD is implemented using the SigLIP2 series as vision encoder, Qwen2.5 family as LLM and a 2-layer MLP as projector. For the Teacher, we consistently use a 384×384 vision encoder, which produces 729 visual tokens. For the Student, we train at two resolutions (224×224 and 384×384). All three training stages are performed using bfloat16 numerical precision and DeepSpeed ZeRO-2 optimization across a single academic node equipped with 6 NVIDIA RTX 4090 (24GB) GPUs. Benefiting from our chunked loss implementation and parameter-free CRRP operator, the entire three-stage pipeline (DPT + SFT + CDFT) for the Qwen2.5-0.5B model at 224×224 resolution takes only approximately 13 wall-clock hours in total (with DPT taking ∼4.5h, SFT ∼2.0h, and CDFT

∼6.5h). Peak GPU memory consumption remains strictly under 21 GB across all stages, demonstrating that CM-FDD is highly computationally efficient and readily reproducible on consumer-grade hardware without requiring industrial-scale HPC clusters. Detailed training hyper-parameters are shown in Table 5.

Table 5: Training configurations and hyperparameters for DPT, SFT, and CDFT.

	DPT	**SFT**	**CDFT**
Trainable Modules			
Vision Encoder	×	×	×
Projector	✓	✓	✓
LLM	×	✓	✓
Hyperparameters			
LLM sequence length	2048	2048	2048
Optimizer	AdamW	AdamW	AdamW
Optimizer hyper-parameter	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$
Learning rate	5e-4	1e-5	3e-6
Connector learning rate	–	2e-6	5e-7
Learning rate scheduler	Cosine decay	Cosine decay	Cosine decay
Weight decay	0.0	0.0	0.0
Training epoch	1.0	1.0	1.0
Warm-up ratio	0.03	0.03	0.03
Global batch size	48	384	48
Numerical precision	bfloat16	bfloat16	bfloat16
Model parallelism	DeepSpeed ZeRO-2	DeepSpeed ZeRO-2	DeepSpeed ZeRO-2

Next, we elaborate on the key hyperparameters specific to each stage. First, in the DPT stage, λ_{dpt} is set to 100.0 to scale the state loss to an adequate supervision magnitude. Then, in the CDFT stage, $\lambda_{\text{dyn}}^{\max}$ and $\lambda_{\text{logit}}^{\max}$ are set to 0.3 and 0.65, respectively. β, τ_{aff} and τ_{logit} are set to 0.05, 0.2 and 2.0, respectively. To reduce peak memory usage, we compute the language-modeling loss in chunks without materializing logits for the full sequence. In CDFT, we use chunks of 512 tokens for cross-entropy and 256 valid token positions for logit-KL distillation.

CM-FDD supervises the cross-modal affinity states at a small number of matched depths rather than at every layer, and the two stages that use this supervision select those depths differently.

DPT. In the first stage, the student LLM is frozen and only the projector is optimized. The grounding-state term therefore only needs to shape the visual interface feeding the frozen network. To this end, we supervise a sparse subset $\mathcal{J}_{\text{DPT}}$ consisting of three depths: an early, a middle, and a late depth. We use the same triple for both students, matching student layers $\{2, 9, 23\}$ with teacher layers $\{3, 10, 27\}$, and Eq. 12 averages $\mathcal{L}_{\text{state}}$ over this subset. Since only the projector is trainable, three depths suffice to ensure that the projected visual features induce teacher-compatible grounding across the entire frozen LLM, covering both early multimodal fusion and late semantic refinement.

CDFT. In the third stage both the student LLM and the projector are trained and the full dynamics loss is applied, so we set the number of pairs K according to the depth of the student instead of sharing a single value. For the 24-layer Qwen2.5-0.5B student we use K=4 equally spaced depths, $\{4, 9, 14, 19\}$, matched with $\{5, 10, 16, 27\}$ of the 36-layer Qwen2.5-3B teacher. For the 28-layer Qwen2.5-1.5B student we use K=6 non-equidistant depths, $\{2, 5, 9, 14, 19, 23\}$, matched with $\{3, 6, 10, 15, 21, 27\}$. The sampling of the 1.5B student is deliberately non-uniform: three of its six pairs (student layers 2, 5 and 9) lie in the first third of the network, where text-to-vision affinities are initially formed, while the spacing widens in the middle of the network, where the grounding pattern evolves most smoothly; on the teacher side the pairs are likewise denser in the early and middle layers (consecutive gaps of 3, 4, 5, 6 and 6 layers), and the deepest pair of both students is placed in the last quarter of the network (relative depth 0.83 and 0.85, respectively) so that the consolidated grounding state is supervised as well. This configuration lets the student inherit the teacher's

cross-modal reasoning dynamics across the full depth of the network, with particular emphasis on the critical phases where text-to-vision affinities are first formed and subsequently consolidated.

C Teacher Models: Training, Evaluation, and Distillation Behavior

Both teachers pair a Qwen2.5 language model with a frozen vision encoder at 384px and are trained on an identical two-stage LLaVA recipe with DeepSpeed ZeRO-2 in bf16 and a 2,048-token context. The alignment stage trains only the two-layer MLP connector on the 558K image–caption subset, leaving both the LLM and the vision tower frozen; the instruction-tuning stage then trains the connector jointly with LoRA adapters inserted into every linear projection of the LLM (query, key, value, output, gate, up and down), again with the vision tower frozen, on the 665K LLaVA-1.5 mixture. We use rank 256 with $\alpha = 512$ for the 3B teacher, corresponding to 485M trainable parameters (12.1% of the model), and rank 128 with $\alpha = 256$ for the 7B teacher (663M trainable parameters, 7.6%), both with dropout 0.05 and no bias terms. Each stage runs for a single epoch under a cosine schedule with a linear warm-up over the first 3% of updates and a peak learning rate of 1×10^{-4}; instruction tuning comprises 6,926 updates at an effective batch size of 96 examples and takes 12.7h for the 3B and 18.5h for the 7B teacher, reaching final training losses of 0.85 and 0.71.

In the main experiments, the teacher model serves only as a supervision signal and is not deployed at inference time. To make the capacity and behavior of the teachers explicit, we additionally evaluate both the 3B and 7B teachers on the same benchmark suite used for the students. These results are reported in Table 6.

Table 6: Evaluation results of the 3B and 7B teachers on the same benchmark suite. These models are used only during training and are not deployed at inference time.

Teachers	VQAv2	GQA	VizWiz	SQA	TextVQA	MME	MMB	MMB_{CN}	POPE	MMMU	Avg_7	Avg_{10}
Qwen2.5-3B	**81.4**	**64.1**	53.8	76.9	63.7	93.7	81.6	79.2	88.6	41.0	73.3	72.4
Qwen2.5-7B	81.3	62.6	**59.6**	**81.4**	**66.2**	**96.1**	**82.6**	**81.0**	**88.8**	**44.7**	**75.6**	**74.4**

Table 6 shows that the 7B teacher is stronger in aggregate, improving Avg_7 from 73.3 to 75.6 and Avg_{10} from 72.4 to 74.4 over the 3B teacher. The gains are broad, with notable improvements on VizWiz (+5.8), SQA (+4.5), MMMU (+3.7), MMB_{CN} (+1.8), and TextVQA (+2.5), while the 3B teacher is slightly better only on VQAv2 (+0.1) and GQA (+1.5). Nevertheless, the main distillation results show that the 3B teacher yields a more effective student than the 7B teacher. This is consistent with LLaVA-KD (Cai et al., 2025), which reports the same counterintuitive pattern: larger teachers do not necessarily produce better students. In their Table 8, Qwen2.5-3B as teacher gives 58.9 Avg_{10} with the DPT-SFT-DFT recipe, whereas Qwen2.5-7B as teacher gives 58.3, despite the 7B teacher being stronger on its own benchmarks (69.3 vs. 66.2 Avg_{10} in their evaluation). LLaVA-KD attributes this to the teacher–student capacity gap: beyond a certain point, a stronger teacher creates an alignment bottleneck rather than better supervision. Our results follow the same trend, suggesting that the 3B teacher is a better-matched supervision source for compact students, while the 7B teacher's additional capacity does not translate into more transferable cross-modal grounding dynamics. We therefore use the 3B teacher in the main experiments and report the 7B teacher here only for transparency.

D Supplementary Ablation Study

Here we provide additional ablation experiments to verify the rationality and robustness of our design.

D.1 All-Token Affinity Supervision

The affinity loss considered in this section averages over all valid non-visual text tokens, including system, user, and assistant tokens. Let Ω_b^{all} denote the valid non-visual text-token positions of image-

containing sample b, i.e. the non-padding, non-special, and non-image-placeholder positions, and let $\mathcal{B}_I$ denote the image-containing samples in a mini-batch. The corresponding state loss is

$$\mathcal{L}_{\text{state_all}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sum_{b\in\mathcal{B}_I}|\Omega_b^{\text{all}}|}\sum_{b\in\mathcal{B}_I}\sum_{n\in\Omega_b^{\text{all}}} D_{\text{KL}}\big(\mathbf{A}_{T,k}[b,n,:]\,\big\|\,\mathbf{A}_{S,k}[b,n,:]\big)\,, \tag{33}$$

where K is the number of matched layers. This formulation provides broad contextual supervision over the complete valid textual context. Because the normalization is performed over all valid affinity rows in the mini-batch, each valid text position contributes with equal weight, so samples with longer textual contexts carry proportionally more weight in the objective than samples with shorter contexts. As shown in Table 7, Avg_7 decreases by 0.7 and Avg_{10} by 0.3. We therefore retain the answer-focused variant that retains only valid assistant-response positions as the main method, since it performs better than the text–vision relations from the full instruction.

Table 7: Effect of all-token affinity supervision. Results use the Qwen2.5-0.5B student at resolution 224.

Method	VQAv2	GQA	VizWiz	SQA	TextVQA	MME	MMB	MMB_{CN}	POPE	MMMU	Avg_7	Avg_{10}
CM-FDD	**71.1**	**56.2**	**42.0**	**60.1**	45.0	**72.2**	**66.4**	**62.9**	**85.7**	32.2	**57.8**	**59.4**
- Answer-token balanced	71.0	56.1	38.0	60.0	**45.1**	72.0	66.2	62.5	**85.8**	**33.8**	57.1	59.1

D.2 Sample-Level Affinity-Derivative Cosine Loss

The main method aligns the layer-wise change in affinity distributions using the stabilized per-row cosine loss in Eq. 8. Here we evaluate a sample-level variant that instead aggregates all valid assistant-response-token and visual-token entries within each sample before computing the cosine distance. For each transition k and sample $b \in \mathcal{B}_k$, we flatten the affinity change into a single vector,

$$\mathbf{d}_{m,k,b} = \text{vec}(\Delta\mathbf{A}_{m,k,b}) \in \mathbb{R}^{|\mathcal{V}_b|N_S^{\text{vis}}}, \qquad m \in \{S,T\}.$$

The sample-level evolution loss is then

$$\mathcal{L}_{\text{evo_sample}} = \frac{1}{K-1}\sum_{k=2}^{K}\frac{1}{|\mathcal{B}_k|}\sum_{b\in\mathcal{B}_k}\left[1-\frac{\mathbf{d}_{T,k,b}^{\top}\mathbf{d}_{S,k,b}}{\max(\|\mathbf{d}_{T,k,b}\|_2\|\mathbf{d}_{S,k,b}\|_2,\epsilon)}\right], \qquad \epsilon = 10^{-8}. \tag{34}$$

This variant produces one cosine similarity per sample and then averages over valid samples, so that samples with longer responses do not contribute more terms than shorter ones. It does not explicitly detect rows in which only one model changes while the other remains stationary, since such degenerate transitions are absorbed into the sample-level vector. Table 8 shows that replacing the stabilized per-row treatment with this sample-level aggregation reduces aggregate performance: Avg_7 and Avg_{10} decrease by 0.2 and 0.1, respectively. We therefore retain the stabilized per-row loss in the main method.

Table 8: Effect of replacing the stabilized per-row affinity-derivative loss with a sample-level variant. Results use the Qwen2.5-0.5B student at resolution 224.

Method	VQAv2	GQA	VizWiz	SQA	TextVQA	MME	MMB	MMB_{CN}	POPE	MMMU	Avg_7	Avg_{10}
CM-FDD	**71.1**	**56.2**	**42.0**	60.1	45.0	**72.2**	**66.4**	**62.9**	**85.7**	32.2	**57.8**	**59.4**
- Sample-level cosine	71.0	56.2	41.0	**60.2**	44.9	72.0	66.4	62.6	85.6	**32.8**	57.6	59.3

D.3 Ablation Study on Training Recipes

As shown in Table 9, the proposed DPT+SFT+CDFT schedule achieves the best results, reaching 57.8 Avg_7 and 59.4 Avg_{10}. Adopting CDFT after SFT improves Avg_7/Avg_{10} by 2.7/1.9 points over DPT+CDFT+SFT, whereas performing CDFT before SFT is markedly less effective. This confirms that DPT should initialize visual-language alignment, SFT should first establish instruction following, and CDFT should subsequently refine cross-modal reasoning.

Table 9: Training-recipe ablation.

Recipe	Avg_7	Avg_{10}
DPT+CDFT	52.3	54.8
DPT+SFT	56.2	58.0
DPT+CDFT+SFT	55.1	57.5
DPT+SFT+CDFT	**57.8**	**59.4**

D.4 Effect of the LLM Backbone

Setup. Our main experiments instantiate CM-FDD within the Qwen2.5 family. To verify that the framework is not tied to this choice, we repeat the complete pipeline with a different LLM family while keeping every other component fixed. The student is **MobileLLaMA-1.4B-Base** equipped with the vision encoder at 224×224 resolution and a two-layer MLP projector; the teacher is **MobileLLaMA-2.7B-Base** with the higher-resolution 384×384 vision encoder and the same projector design. The teacher is first aligned to the visual modality by training only its projector on LLaVA1.5-558k, and is then fine-tuned on LLaVA-mix-665k with LoRA (r=256, α=512, dropout 0.05) while the projector continues to be optimized at a $10\times$ smaller learning rate. The student then follows the three stages of the main configuration, DPT, SFT and CDFT, on the same data and with the same loss composition, optimizer, schedule and DeepSpeed ZeRO-2 setup under bf16. Layer pairs are selected by the same relative-depth rule with K=6 pairs and the same student indices $\{2, 5, 9, 14, 19, 23\}$ ($\{2, 9, 23\}$ for DPT); because the MobileLLaMA teacher has 32 layers instead of the 36 layers of the Qwen2.5-3B teacher, the matched teacher indices become $\{3, 7, 12, 19, 26, 31\}$ ($\{3, 12, 31\}$ for DPT).

Results. Apart from the peak learning rates of the SFT and CDFT stages (8×10^{-6} and 2×10^{-6}, with projector rates 2×10^{-6} and 2×10^{-7}; DPT keeps 5×10^{-4} with λ_{dpt}=100), every hyperparameter is identical to the main configuration. Table 10 compares the two backbones.

Table 10: Effect of the LLM backbone. The last row is the main configuration (Qwen2.5-1.5B student distilled from a Qwen2.5-3B teacher); the row above it replaces both models with the MobileLLaMA family.

Method	LLM	Res.	VQAv2	GQA	VizWiz	SQA	TextVQA	MME	MMB	MMB_{CN}	POPE	MMMU	Avg_7	Avg_{10}
LLaVADI	MobileLLaMA-2.7B	336	–	**61.4**	–	64.1	50.7	68.8	62.5	–	86.7	–	–	–
MOVE-KD	MobileLLaMA-1.4B	336	71.5	57.7	28.6	56.0	42.6	57.3	47.0	–	84.3	–	–	–
Beta-KD†	MobileLLaMA-1.4B	336	**76.9**	61.2	44.5	62.9	**54.2**	67.5	60.2	49.8	86.0	30.7	57.2	59.4
CM-FDD	MobileLLaMA-1.4B	224	73.1	57.1	30.5	54.1	44.2	72.2	65.7	44.4	86.5	29.0	52.6	55.7
CM-FDD	MobileLLaMA-1.4B	384	75.4	58.7	31.6	54.5	50.8	69.1	67.8	48.0	**87.9**	28.6	54.4	57.2
CM-FDD	Qwen2.5-1.5B	224	76.6	61.0	**51.2**	**73.2**	54.1	**82.3**	**77.4**	**74.5**	86.6	**35.8**	**67.7**	**67.3**

CM-FDD applies to the MobileLLaMA pair without any modification of the method, and object-level grounding transfers almost losslessly: POPE is on par with the Qwen student (86.5 vs. 86.6) and VQAv2/GQA stay within 3.5/3.9 points. The remaining gap is concentrated in benchmarks that depend on the student's own language and world knowledge rather than on cross-modal grounding: $\text{MMBench}_{\text{CN}}$ (-30.1), VizWiz (-20.7), ScienceQA (-19.1), MMBench (-11.7), MME (-10.1), TextVQA (-9.9) and MMMU (-6.8), which accumulates into -15.1 Avg_7 and -11.6 Avg_{10}. Since the perception budget (224×224, 256 tokens) and the distillation losses are identical in both rows, this pattern is attributable to the LLM priors: Qwen2.5-1.5B is pre-trained on a much larger and multilingual corpus, whereas MobileLLaMA-1.4B-Base is English-centric. The two extremes are consistent with this reading: the abilities that mainly require re-locating the visual evidence provided by the teacher transfer across family, while the abilities that additionally require encyclopedic, OCR or Chinese knowledge remain bounded by the pre-training of the student LLM. In short, CM-FDD transfers cross-modal dynamics across LLM families, but the absolute performance ceiling of the student is still set by its base LLM. We note that this ablation replaces the LLM on both sides of the distillation, and that the two teachers have comparable capacity (2.7B vs. 3B), so the comparison measures the backbone choice as a whole rather than the student LLM in isolation.

D.5 Statistical Robustness Across Training Seeds

To assess the robustness of CM-FDD to training randomness, we repeat the complete DPT–SFT–CDFT pipeline independently using seeds 1234, 42, 43, and 44. For each metric, we report the mean and sample standard deviation across the four runs. We compute two-sided 95% Student-t confidence intervals as

$$\bar{x} \pm t_{0.975,\,n-1} \frac{s}{\sqrt{n}}, \qquad n = 4, \tag{35}$$

where $\bar{x}$ and s denote the sample mean and standard deviation, respectively, and $t_{0.975,3} = 3.182446$. As shown in Table 11, CM-FDD exhibits stable performance across independent runs. In particular,

Avg_7 reaches 57.8081 ± 0.0915 and Avg_{10} reaches 59.3087 ± 0.1984, indicating that the reported gains are not attributable to a single favorable random seed.

Table 11: Results over four independent DPT–SFT–CDFT runs. Intervals are two-sided 95% Student-t confidence intervals.

Metric	Avg_{10}	Avg_7	VQAv2	GQA	VizWiz	ScienceQA	TextVQA	MME	MMBench	$\text{MMBench}_{\text{CN}}$	POPE	MMMU
Mean	59.3087	57.8081	71.0762	55.9072	42.5665	60.1214	44.5050	72.8890	66.2220	62.4452	85.3042	32.0500
Std.	0.1247	0.0575	0.0184	0.1810	1.0713	0.1557	0.4611	0.9768	0.3225	0.3872	0.4601	0.6245
95% CI	±0.1984	±0.0915	±0.0292	±0.1810	±1.7044	±0.2477	±0.7337	±1.5541	±0.5131	±0.6160	±0.7320	±0.9936

E Additional Analysis of Visual Token Downsampling Methods

When the teacher and student vision encoders operate at different resolutions, the teacher produces $27 \times 27 = 729$ visual tokens, whereas the student uses $16 \times 16 = 256$ tokens. Our main method applies bilinear interpolation to the teacher feature map. Given $H_T^{vis} \in \mathbb{R}^{B\times 729\times D_T}$, we reshape the tokens into a 27×27 spatial grid, interpolate each feature channel, and flatten the result:

$$\widetilde{H}_T^{vis} = \text{Flatten}\big(\text{Bilinear}_{27\to 16}\left(\text{Grid}_{27\times 27}(H_T^{vis})\right)\big) \in \mathbb{R}^{B\times 256\times D_T}. \tag{36}$$

Bilinear interpolation is parameter-free and preserves the spatial ordering of the visual tokens while combining nearby teacher patches with distance-dependent weights.

We compare bilinear interpolation with three alternative downsampling methods: the Lightweight Refinement Module (LRM), adaptive average pooling, and Hungarian matching based on visual-feature similarity.

Lightweight Refinement Module. In the implementation, the LRM first performs the same bilinear projection as the main method and then applies a trainable residual depthwise convolution on the resulting 16×16 feature map. Let

$$X_T = \text{Grid}_{16\times 16}\left(\text{Bilinear}_{27\to 16}(H_T^{vis})\right).$$

The LRM computes

$$\widetilde{X}_T = X_T + \text{DWConv}_{3\times 3}(X_T), \qquad \widetilde{H}_T^{vis} = \text{Flatten}(\widetilde{X}_T). \tag{37}$$

The convolution has D_T input and output channels, uses D_T groups, and therefore applies an independent spatial filter to each feature channel. It uses zero padding to preserve the 16×16 spatial resolution. The convolution weights and bias are initialized to zero, so the LRM is exactly equivalent to bilinear interpolation at initialization and learns only a residual correction during training. For a 3×3 kernel, the module contains $10D_T$ trainable parameters, including the bias.

The LRM obtains a score of 57.0, compared with 59.4 for bilinear interpolation. Although the residual convolution can adapt the projected features, its correction is content-independent and is shared across all images. Moreover, the refinement is learned indirectly through the distillation objectives, rather than directly supervised as a spatial reconstruction. The learned correction can consequently disturb the smooth local correspondence provided by bilinear interpolation. Its lower performance suggests that the additional flexibility does not compensate for the spatial prior already encoded by the parameter-free projection.

Average Pooling. For adaptive average pooling, the teacher features are reshaped into a 27×27 grid and pooled to 16×16:

$$\widetilde{X}_T[u, v, :] = \frac{1}{|\mathcal{R}_{u,v}|} \sum_{(i,j)\in\mathcal{R}_{u,v}} X_T[i, j, :], \qquad \widetilde{H}_T^{vis} = \text{Flatten}(\widetilde{X}_T), \tag{38}$$

where $\mathcal{R}_{u,v}$ is the adaptive pooling region associated with output position (u, v). Unlike bilinear interpolation, average pooling assigns equal weight to the patches in each pooling region. It therefore preserves the coarse spatial layout but may blur small objects, boundaries, and discriminative attributes that are important for visual grounding.

Average pooling obtains 58.2, which is below the 59.4 achieved by bilinear interpolation. This difference indicates that distance-aware interpolation weights provide a more informative local correspondence than uniform regional averaging for cross-modal affinity distillation.

Hungarian Matching Based on Visual Features. The implementation referred to as Hungarian matching constructs the assignment from vision encoder features rather than language-model logits. Let $Z_{T,i}$ and $Z_{S,q}$ denote the teacher and student visual features. For every teacher token i and student token q, it computes

$$C_{iq} = \cos(Z_{T,i}, Z_{S,q}) = \frac{Z_{T,i}^{\top} Z_{S,q}}{\|Z_{T,i}\|_2 \|Z_{S,q}\|_2}. \tag{39}$$

The Hungarian algorithm then finds a one-to-one assignment $\pi^\star$ maximizing the total cosine similarity:

$$\pi^\star = \arg\max_{\pi} \sum_{q=1}^{256} C_{\pi(q),q}, \qquad \widetilde{H}_T^{vis}[q,:] = H_T^{vis}[\pi^\star(q),:]. \tag{40}$$

Thus, each student visual position receives one distinct teacher token, while the remaining $729 - 256 = 473$ teacher tokens are discarded. The matching indices are computed from detached features and are not learned through the distillation loss.

Hungarian matching obtains 58.2. Its hard one-to-one assignment can select semantically similar but spatially distant teacher patches, because the matching objective does not impose a neighborhood or grid-topology constraint. It also discards a substantial portion of the teacher representation and produces discontinuous correspondences between adjacent student positions. In contrast, bilinear interpolation uses information from all teacher patches and preserves local spatial smoothness. These properties make bilinear interpolation better suited to constructing stable cross-modal affinity targets.

Overall, bilinear interpolation achieves the best performance, with a score of 59.4, while LRM, average pooling, and Hungarian matching obtain 57.0, 58.2, and 58.2, respectively. The results suggest that the effectiveness of visual-token alignment depends not only on feature similarity, but also on preserving the spatial topology and local continuity of the original teacher feature map. Bilinear interpolation provides this inductive bias without introducing additional trainable parameters or hard, discontinuous token selection.

F Diagnostic Probes of Grounding Transfer

CM-FDD is defined on the dynamics of the cross-modal affinity map, which describes how the distribution over patches that supports each answer token is reallocated from one depth to the next, rather than on any single-layer snapshot of it. The two probes collected here test that formulation from opposite ends, and are constructed so that neither can substitute for the other.

The first probe is *representational*. It asks whether the quantity on which the objective is defined is the one that actually changes under optimization. To test this, consider two students with identical backbones, vision encoders, and training data, differing only in the presence of the distillation term. Does the distilled student's depth-wise affinity rearrangement track the teacher's more closely than the supervised baseline does? Agreement on such a diagnostic is necessary but not sufficient, since a student may match an internal statistic without that statistic doing any work.

The second probe is therefore *causal*. It intervenes on the vision stream of the trained student at inference time and measures how much the answer suffers. By deleting the patches that the student's affinity map ranks highest, deleting a size-matched random subset, and deleting the lowest-ranked patches, we obtain three monotonically ordered damage curves, while global controls that remove or scramble the entire image provide an upper bound on the effect. Taken together, the two probes connect the training signal to an internal signature (§F.1) and this signature to externally observable behaviour (§F.2).

F.1 Depth-Wise Affinity Dynamics Converge on the Teacher

We instantiate both students from the same Qwen2.5-1.5B backbone and the same vision encoder, and train them on the identical LLaVA-1.5 665K mixture; the only difference is the objective (`sft` versus `cdft`). Both are compared against the Qwen2.5-3B teacher that supplies the distillation targets, on the same 256 examples. For each example the probe extracts the answer-token-to-patch

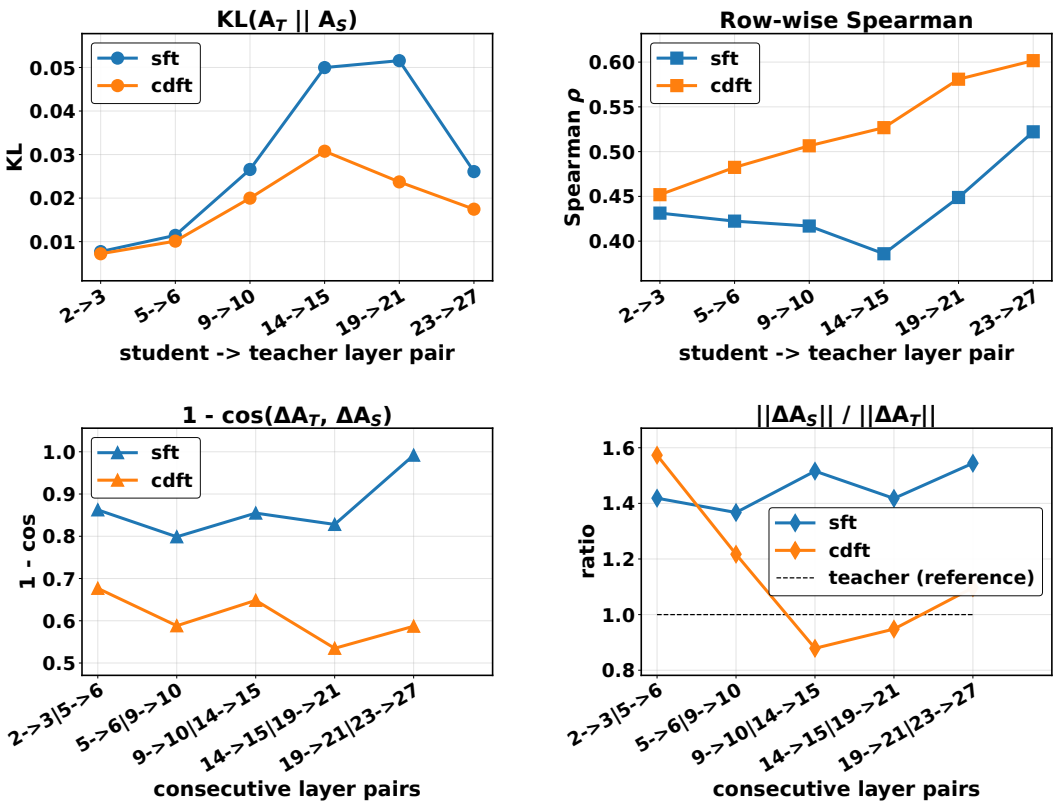


Figure 7: Depth-wise affinity dynamics of two 1.5B students trained on the identical 665K mixture, evaluated against the 3B teacher over 256 examples. **Top row:** per-layer agreement between the student and teacher affinity maps, measured by KL divergence (left) and row-wise Spearman correlation (right). **Bottom row:** agreement between the inter-layer rearrangements ΔA, measured by $1 - \cos(\Delta\mathbf{A}_T, \Delta\mathbf{A}_S)$ (left) and by the ratio of rearrangement magnitudes (right; the dashed line marks the teacher's own magnitude). Error bars are 95% bootstrap intervals. The distilled student lies closer to the teacher on all four readouts, and its advantage is concentrated in the middle and late layers, where the supervised baseline both overshoots the teacher's rearrangement magnitude and points in a nearly orthogonal direction.

affinity map $\mathbf{A}_{m,k}$ at six student depths and at the six teacher depths they are paired with ($2 \to 3$, $5 \to 6$, $9 \to 10$, $14 \to 15$, $19 \to 21$, $23 \to 27$). Because the teacher operates on a 384px grid, its patch grid is bilinearly resampled onto the student's 256 tokens so that both maps live on the same support; affinities on both sides use the same temperature ($\tau = 0.2$) and the same answer-token masking. Two families of readouts follow. Per-layer agreement compares the maps themselves: the KL divergence $\mathrm{KL}(\mathbf{A}_T \,\|\, \mathbf{A}_S)$ and the row-wise Spearman correlation, which measures whether the teacher's patch *ranking* is recovered rather than merely its magnitudes. Per-pair agreement compares the rearrangement $\Delta\mathbf{A}_{m,k} = \mathbf{A}_{m,k} - \mathbf{A}_{m,k-1}$, along two independent axes: its direction, $1 - \cos(\Delta A_T, \Delta A_S)$, and its scale, $\|\Delta\mathbf{A}_S\|/\|\Delta\mathbf{A}_T\|$, for which unity is the teacher's own magnitude.

Three observations emerge from Fig. 7. First, the KL divergence is lower for the distilled student at every depth, with a mean of 0.0182 against 0.0289 (-37%); the largest absolute gaps lie in the middle of the network, where KL drops from 0.0266 to 0.0200 at $9 \to 10$ and from 0.0500 to 0.0308 at $14 \to 15$, so the improvement is concentrated where the visual evidence is progressively converted into answer-relevant structure rather than at the embedding-adjacent layers. Second, the ranking readout moves the same way, from a mean Spearman ρ of 0.438 to 0.525, with the two largest gains ($+0.141$ and $+0.132$) again in the middle and late layer pairs; the distilled student therefore orders the patches more like the teacher, not merely scores them on a similar scale. Third, and most diagnostic, the direction of the rearrangement: the supervised baseline is almost orthogonal to the teacher at $19 \to 21$, with $1 - \cos = 0.992$ (95% CI $[0.983, 1.001]$), whereas the distilled student

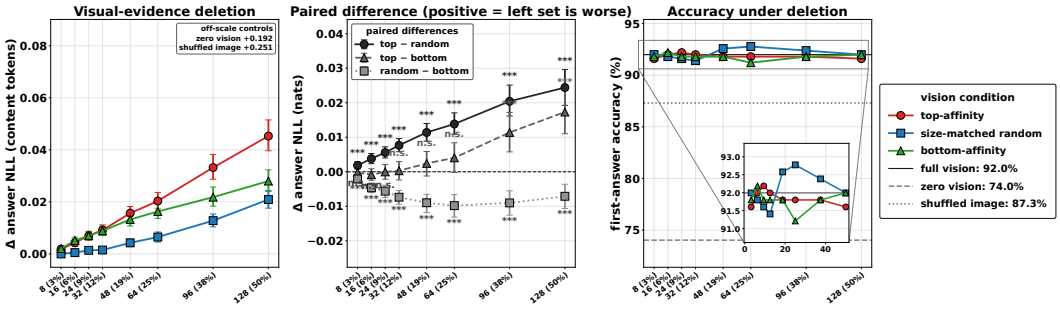


Figure 8: Partial deletion of the vision stream for the 0.5B CDFT student (512 examples, in-distribution blur, answer-content NLL). **Left:** answer NLL damage as a function of the number of deleted patches, annotated with the corresponding share of the 256 vision tokens, for the top-affinity, size-matched random and bottom-affinity sets ranked at the deepest distilled pair; the off-scale box reports the two global controls. **Middle:** the paired differences behind these curves, positive meaning that the left set is worse; error bars are 95% paired-bootstrap intervals and stars denote two-sided paired-bootstrap significance ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, n.s. otherwise, with the p-value floored at $1/2000$ so that *** is the finest level this test resolves). **Right:** first-answer accuracy over the same budgets, with the full-vision, zero-vision and shuffled-image levels as horizontal references. Deleting affinity-ranked evidence is monotonically harmful and significantly worse than size-matched random deletion at every budget, yet accuracy stays within 0.4 pp of the full-vision level even when half of the vision tokens are removed.

reaches 0.587 ($[0.579, 0.596]$), and the two intervals are disjoint at every depth. The magnitude ratio explains why this is not a matter of scale: SFT overshoots the teacher by 42–54% at every pair, moving the affinity distribution much further than the teacher does and in a largely unrelated direction, while the distilled student straddles unity (0.88–1.57, mean 1.14). The two students share their backbone, their data and their visual encoder, so the difference isolates the effect of the objective—and that effect is a correction of the *direction* of the depth-wise rearrangement alongside a correction of its size, which is precisely the quantity the formulation targets. Being a descriptive comparison, however, this probe shows the student moving in the teacher's direction without establishing that the movement is what the answers rely on; the deletion probe of the next subsection supplies that half of the argument.

F.2 Causal Necessity and Redundancy of the Transferred Grounding

Agreement on an internal diagnostic is only meaningful if the resulting grounding is actually consulted when the student answers. The second probe therefore intervenes directly on the vision stream of the CDFT student at inference time, under the identical generation protocol, and measures the damage to the answer.

F.2.1 Partial Deletion of Affinity-Ranked Evidence

At each budget $k \in \{8, 16, 24, 32, 48, 64, 96, 128\}$ patches from 3% to 50% of the 256 vision tokens, the selected patch embeddings are replaced by the 5×5 local mean of the patch-feature grid, the cached-feature analogue of a pixel-space Gaussian blur, so that the model is shown a locally uninformative but in-distribution view rather than an out-of-distribution token. Three sets of equal size are compared: the k highest-affinity patches of the deepest distilled pair ($19 \to 27$), a uniformly random subset of the same size, and the k lowest-affinity patches. Damage is the increase in the negative log-likelihood of the gold answer, averaged over answer-content tokens only, which excludes punctuation and the fifty most frequent answer templates so that formatting tokens, carrying no visual information, do not dilute a token-averaged readout. All contrasts are two-sided paired bootstraps over the 512 examples with 2000 resamples, which cancel per-example difficulty and are what make the top-versus-random gap resolvable at this sample size.

The resulting curves (Fig. 8) are monotone and steepen with the budget. At $k = 128$ the top-affinity set costs $+0.0453$ nats against $+0.0209$ for the size-matched random set and $+0.0280$ for

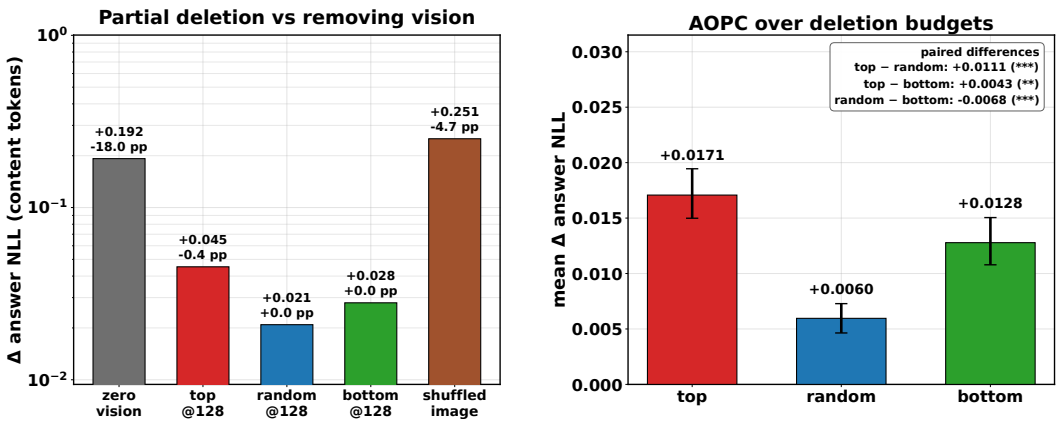


Figure 9: Global vision controls and budget-averaged damage for the same 0.5B CDFT student. **Left:** answer NLL damage (log scale) of removing the vision stream entirely, of replacing the image by a different one, and of deleting the largest patch budget, each bar annotated with the change in first-answer accuracy relative to full vision. **Right:** area over the deletion curve (AOPC, mean damage across budgets) for each of the three deletion sets, with the paired-bootstrap significance of the contrasts marked as before. Partial deletion of the top-affinity set accounts for roughly a quarter of the damage caused by removing vision altogether, the three deletion sets are ordered top $>$ bottom $>$ random, and the two global controls sit an order of magnitude above every partial-deletion budget.

the bottom-affinity set; the paired contrast is $+0.0244$ nats for top minus random ($p < 0.001$, the top set being worse on 70.9% of examples) and $+0.0173$ nats for top minus bottom ($p < 0.001$). Top exceeds random at every one of the eight budgets, and budgeting over the curve yields the area over the deletion curve (AOPC) of 0.0171 for the top-affinity set against 0.0060 for random ($+0.0111$, $p < 0.001$) and 0.0128 for bottom ($+0.0043$, $p = 0.007$).

Two conclusions follow. The first is causal necessity: the patches that the affinity map selects are the patches the student actually uses, since removing only 3% of the grid already costs measurably more than removing 3% at random, and the gap widens rather than saturates as the intervention grows. The second is redundancy: the same 50% deletion leaves first-answer accuracy at 91.6% against 92.0% for full vision ($p = 0.66$, not significant), so the student degrades gracefully and continues to answer correctly from the surviving evidence. The magnitude places this on an absolute scale: deleting half of the vision tokens by affinity recovers 23.5% ($0.0453/0.1923$) of the damage caused by removing the vision stream entirely, so the transferred grounding is necessary but distributed over far more patches than the answer strictly requires.

F.2.2 Global Vision Controls and Budget-Averaged Damage

The global controls in Fig. 9 bound the effect from above and calibrate its scale. In the zero-vision condition the entire vision stream is replaced by zeros; in the shuffled-image condition the same encoder is fed a different image from the batch, which preserves the statistics and magnitude of the visual features while destroying their correspondence with the question. Both are scored with the same readouts as the deletion curves. Removing vision costs $+0.192$ nats and lowers first-answer accuracy from 92.0% to 74.0% (-18.0 pp, $p < 0.001$), whereas showing the wrong image costs $+0.251$ nats but only -4.7 pp (87.3%, $p < 0.001$): the shuffled control is the more damaging in likelihood and the less damaging in accuracy, since a wrong image misleads the model into a confident but incorrect answer whereas a zeroed stream collapses it into a degenerate "no image" state. Placing the deletion budgets on this scale separates the interventions cleanly, from the largest budget of the top-affinity set (0.0453) through the bottom-affinity (0.0280) and random (0.0209) sets, up to the global controls (0.192 and 0.251).

Two features of these results deserve emphasis, and one of them is a caveat about interpretation. First, the top-affinity set is the most damaging of the three under this protocol, which is what a faithfulness claim requires, whereas the ordering between the two control sets is not protocol-independent. Uniformly thinning the patch grid preserves a globally coherent, unbiased view of the

Table 12: Inference cost on a single RTX 4090 (bf16, greedy, 32-token budget, median of 10 repetitions after 3 warm-up iterations). Students use 256 visual tokens and the teacher 729, hence the different sequence lengths. Request rate is end-to-end and therefore sensitive to the generated length (17, 25 and 32 tokens for the students at b=1); throughput should be read from tok/s.

Checkpoint	b	Seq. len	TTFT (ms)	TPOT (ms)	tok/s	req/s	Peak VRAM (MiB)	KV (MiB)
0.5B-SFT	1	298	30.2	18.59	51.8	3.05	1824	3.7
0.5B-SFT	4	298	34.8	19.26	198.2	11.66	1888	14.8
0.5B-SFT	8	298	45.4	19.21	385.4	22.67	1969	29.5
0.5B-CMFDD	1	298	30.4	18.88	51.1	3.00	1828	3.7
0.5B-CMFDD	4	298	35.2	19.65	194.5	11.44	1893	14.8
0.5B-CMFDD	8	298	45.4	19.13	387.3	22.78	1975	29.5
0.5B-CMFDD-no-der	1	298	29.5	18.61	52.5	2.10	1828	3.8
0.5B-CMFDD-no-der	4	298	34.3	19.27	200.2	9.53	1893	15.0
0.5B-CMFDD-no-der	8	298	45.0	19.09	393.6	18.74	1975	29.9
Teacher-3B	1	771	51.7	27.91	34.9	1.09	6932	28.2
Teacher-3B	4	771	175.1	28.83	119.8	3.74	7207	112.9
Teacher-3B	8	771	344.1	28.61	208.0	6.50	7572	225.8

scene, while the bottom-affinity set is structured and, because the deletion operator here replaces the removed patches by zeros, it discards contiguous evidence outright instead of degrading it towards its local statistics; the size-matched random set is therefore not a no-information reference at large budgets. Consistently, the larger VQAv2 deletion study of Sec. 4.3—6,000 examples, budgets averaged, blurred rather than zeroed masking, and a teacher-derived ranking—recovers top $>$ random $>$ bottom, so the two controls trade places while the top-affinity set remains the most damaging in both settings. We consequently rest the faithfulness claim on the top-versus-random and top-versus-bottom contrasts, both of which are significant, and on no claim as to which of the two control sets is the weaker reference. Second, the fact that every masked condition lands within 0.8 pp of the full-vision accuracy while the NLL damage accumulates monotonically is consistent with the redundancy reading of the deletion curves: the student's reliance on the distilled grounding is real, graded, and spread across many patches, which is the behaviour one expects from a representation that has been shaped by dynamics rather than memorised as a small set of salient regions.

G Inference Cost on Consumer Hardware

We profile the deployment cost of the CM-FDD students on a single consumer-grade RTX 4090 (24 GiB) in bf16, a practical deployment setting. Every checkpoint answers the same question ("What is in this image?") about the same 640×427 image with greedy decoding and a 32-token budget; each (checkpoint, batch size) pair is warmed up for three iterations and then timed over ten repetitions, and we report the median. The three students share the 224-resolution tower (256 visual tokens, 298-token prompt) while the 3B teacher uses the 384-resolution tower (729 visual tokens, 771-token prompt), so the teacher column conflates a $3.8\times$ parameter difference with a $2.6\times$ longer prefill and we use it only as a reference point. Table 12 reports TTFT (image encoder, connector, prefill and the first decode step), TPOT (mean inter-token latency of the remaining steps), throughput, peak allocated VRAM, the KV-cache footprint, and analytic FLOPs (2 FLOPs per MAC, quadratic attention and cache-aware decoding).

Three observations follow. First, the distillation objective is free at inference time: the three students agree to within 3% on every metric, share an identical parameter count (0.924 B), prefill cost (0.52 TFLOP), decode cost (1.01 GFLOP per token) and memory footprint, so CM-FDD changes only how the student is trained. Second, latency at small batch is decode-dominated—at b=1 the student spends 30 ms on TTFT (of which the vision tower contributes 10.7 ms, i.e. 35%) but 91% of its 328 ms request on decoding—and batching is nearly free latency-wise: at b=8 throughput rises from 51 to 387 tok/s ($7.6\times$) while TTFT grows by 15 ms, TPOT stays flat (18.9→19.1 ms) and peak VRAM increases by 147 MiB, leaving the student at 1.9 GiB of a 24 GiB card with a 29.5 MiB KV cache. Third, relative to the teacher the student is $3.8\times$ smaller in weights, needs $6.3\times$ fewer FLOPs per decoded token and $10.7\times$ fewer prefill FLOPs, and lowers TTFT by $1.7\times$ at b=1 and $7.6\times$ at b=8; for a 32-token answer the end-to-end latency is 616 ms versus 917 ms. The parameter

count is moreover dominated by the frozen vision tower (428 M of 924 M), so the student's language backbone is the smaller half of its footprint.

Overall, the accuracy reported in the main paper is obtained at no additional inference cost: CM-FDD adds no parameters, FLOPs or latency at test time, and the teacher enters only as an offline training signal. This is what makes the method attractive for edge deployment, where the 0.5B student serves sub-second answers within 2 GiB of a 24 GiB GPU.

H Qualitative Case Studies

This section complements the aggregate metrics with a qualitative view of generated responses. Each figure shows one case and fixes a single image and instruction, reporting the answers of three systems side by side: the SFT baseline, the CM-FDD student, and the teacher. The two students share their backbone, visual encoder, tokenizer, and training data, and differ only in the distillation objective, so differences between their columns isolate the effect of the proposed dynamics supervision, while the teacher column shows the behaviour the student is asked to inherit. Cases cover VQAv2, POPE, MME, and VizWiz open-ended question answering, as well as hallucination probing, perception and reading, and degraded real-user photographs. They are drawn from the corresponding evaluation splits. Each card echoes the image, instruction, and ground-truth answer where available, reproduces response text verbatim, including disagreements, and adds a second free-form turn requesting a detailed description of the same image.

The examples shown in Fig. 10 – Fig. 15 are illustrative rather than evidential and are not scored: no claim in the main text depends on them, and we draw no statistical conclusion from how often one column appears preferable.

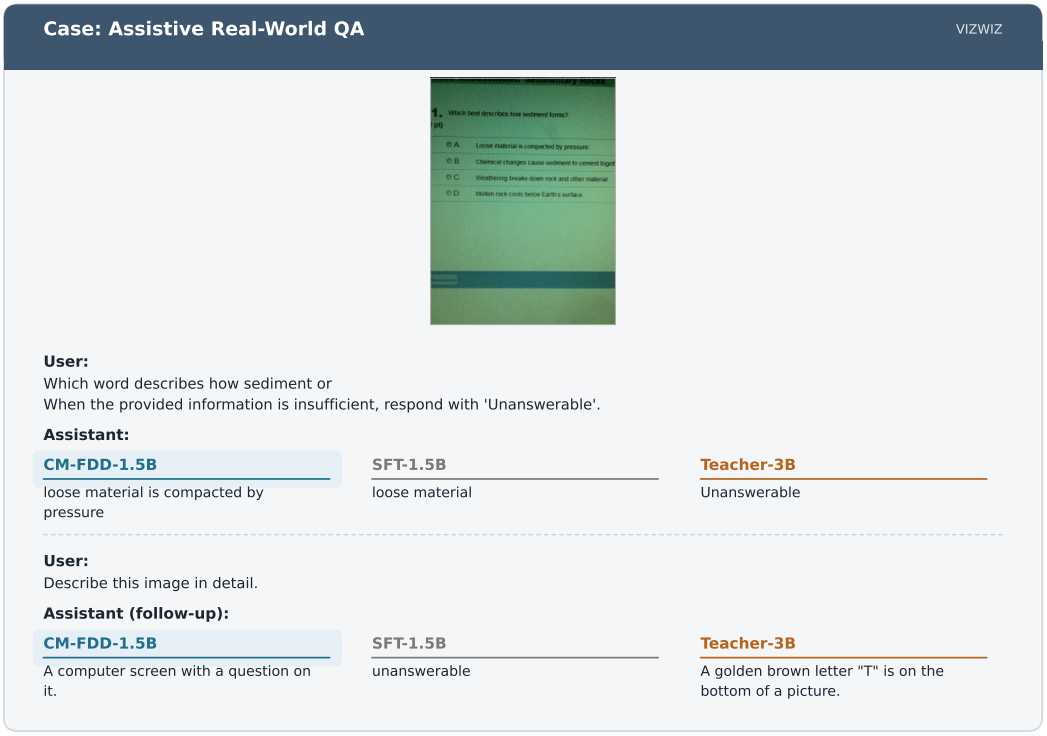


Figure 10: A degraded VizWiz photograph whose question requires the full answer option: CM-FDD grounds its response in the visible material and returns the complete option, whereas SFT-1.5B returns only a fragment and the teacher abstains as *unanswerable*.

Case: Perception — Count MME



User:
Is there a train in the picture?

Assistant:

CM-FDD-1.5B
Yes, there is a train in the picture, specifically a blue and white passenger train.

SFT-1.5B
Yes, there is a train in the picture, and it is a passenger train.

Teacher-3B
No, there is no train in the picture. The image features a single silver-colored object, which is a silver-colored sphere or a silver-colored sphere with a red center.

User:
Describe this image in detail.

Assistant (follow-up):

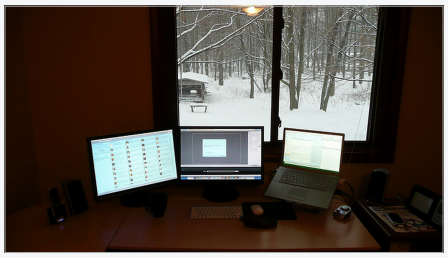
CM-FDD-1.5B
The image features a long blue and white passenger train traveling down the tracks. The train is stopped at a station, and there are several people visible in the scene. Some of them are standing near the train, while others are walking or waiting nearby. The train appears to be a commuter train, as it is likely designed to transport passengers between different locations.

SFT-1.5B
The image features a long passenger train traveling down a track, with its windows open. The train is blue and white, and it appears to be a commuter train. There are several people visible in the image, likely passengers on the train. Some of them are sitting, while others are standing or walking around the train. The train is positioned in the middle of the image, occupying a significant portion of the frame.

Teacher-3B
The image features a single silver-colored object, which is a silver-colored sphere or a silver-colored sphere with a red center. The sphere is the main focus of the image, and it appears to be the only object present.

Figure 11: An MME counting case in which CM-FDD and SFT-1.5B both answer correctly, while the teacher hallucinates an object that is absent from the image.

Case: Visual Question Answering VQAV2 | GT: 3, 2, 1



User:
How many computers is there?

Assistant:

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
There are three computers on the desk.	There are three computers on the desk.	There are no computers in the image. The image shows a single computer monitor, which is the main focus of the scene.

User:
Describe this image in detail.

Assistant (follow-up):

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
The image features a desk with three computers, including two monitors and a laptop. The monitors are placed side by side, with one on the left and the other on the right. The laptop is positioned in the middle of the desk. In addition to the computers, there are a few other items on the desk, such as a cell phone, a cup, and a mouse. The desk is situated in front of a window, allowing natural light to illuminate the workspace. The overall scene suggests a well-equipped and organized workspace for multitasking or working on multiple tasks simultaneously.	The image features a desk with three computers, including a laptop and two monitors. The laptop is placed on the left side of the desk, while the two monitors are positioned on the right side. The desk is also equipped with a keyboard and a mouse, making it a complete workstation. In addition to the computers and peripherals, there are a few other items on the desk, such as a cell phone and a cup. The desk is situated in front of a window, which provides natural light and a view of the outdoors. The overall scene gives off a cozy and functional workspace atmosphere.	The image features a single computer monitor, which is the main focus of the scene. The monitor is placed on a desk, and there is a computer in the background. The desk appears to be a part of a workspace, and the computer monitor is the central element of the image.

Figure 12: A VQAv2 counting question on computers: CM-FDD returns the correct number, in agreement with SFT-1.5B, while the teacher claims the objects are not present.

Case: Visual Question Answering VQAV2 | GT: brown, unknown, black



User:
What color hair does the man have?

Assistant:

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
The man has brown hair.	The man has brown hair.	The man has blonde hair.

User:
Describe this image in detail.

Assistant (follow-up):

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
The image is a vintage photograph of a young man, possibly a boy, wearing a suit and tie. He is wearing a bow tie, which adds a touch of elegance to his appearance. The man is looking directly at the camera, giving the photograph a sense of authenticity and capturing a moment in time. The overall style of the photograph suggests that it was taken during a time when formal attire was more common, and the man's attire reflects that.	The image features a man wearing a suit, standing in front of a camera. He is dressed in a suit with a tie, and his facial features are well-defined. The man appears to be a young man, possibly a boy, and is posing for the camera. The overall composition of the image is focused on the man's attire and his facial expression.	The image features a man with blonde hair, standing in front of a small, square-shaped object. The man is wearing a tie, and he appears to be posing for the camera. The small object in front of him could be a piece of paper, a small toy, or a decorative item.

Figure 13: A VQAv2 attribute question on hair colour: CM-FDD reports the colour supported by the image, in agreement with SFT-1.5B, whereas the teacher answers *blonde*.

Case: Visual Question Answering VQAV2 | GT: 36, 12, 5, 10



User:
How many inch of snow is there?

Assistant:

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
12	12	The image does not provide information about the amount of snow. It only shows a person standing in the snow.

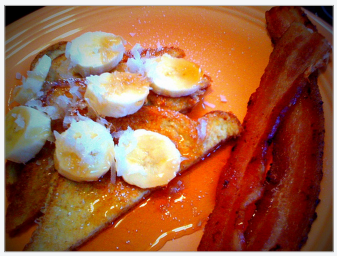
User:
Describe this image in detail.

Assistant (follow-up):

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
The image features a snow-covered mountain range with a person wearing skis and a backpack, standing on a snow-covered slope. The person appears to be taking a break and looking at the beautiful view of the mountains. There are several other people in the scene, some of them also wearing skis and backpacks, indicating that they are likely enjoying the snowy mountain environment as well. The snow-covered slope is quite large, stretching from the foreground to the background, and the mountains in the distance provide a stunning backdrop for the skiers. The scene captures the essence of winter sports and outdoor activities in a picturesque setting.	The image features a snowy mountain landscape with a person wearing skis standing on the snow-covered ground. The skier is positioned in the middle of the scene, with the snow-covered mountains in the background. There are several other people in the scene, some of them also wearing skis, and they are scattered throughout the area. In addition to the skiers, there are a few backpacks visible in the scene, likely belonging to the people enjoying the snowy mountain. The overall atmosphere of the image is one of winter sports and outdoor recreation.	The image features a person standing in the snow, possibly on a street or in a park. The person appears to be wearing a coat, indicating that the weather is cold. The snow is covering the ground, and the person is likely enjoying the winter scenery.

Figure 14: A VQAv2 counting question: CM-FDD correctly answers “12”, in agreement with SFT-1.5B (also “12”), while the teacher declines to answer, stating that the image does not provide information about the amount of snow.

Case: Object Presence Probing — Adversarial POPE



User:
Is there a banana in the image?

Assistant:

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
Yes, there is a banana in the image, and it is placed on top of the toast.	Yes, there is a banana in the image, and it is placed on top of the toast.	No, there is no banana in the image. The image features a banana peel on a table.

User:
Describe this image in detail.

Assistant (follow-up):

CM-FDD-1.5B	SFT-1.5B	Teacher-3B
The image features a plate with a delicious breakfast meal consisting of toast and a banana. The toast is topped with a generous amount of syrup, which adds a sweet and flavorful touch to the meal. The banana is placed on top of the toast, adding a touch of natural sweetness and a unique texture to the dish. The combination of the banana and syrup on the toast creates a delightful and satisfying breakfast experience.	The image features a plate with a delicious breakfast meal consisting of toast and a banana. The toast is topped with a generous amount of syrup, which adds a sweet and flavorful touch to the dish. The banana is placed on top of the toast, creating a visually appealing and appetizing combination. The plate is placed on a table, ready to be enjoyed by someone who is ready to start their day with a delicious breakfast.	The image is a close-up of a banana peel on a table. The peel is placed on a surface, possibly a table or a countertop, and it appears to be the only object in the scene. The focus of the image is on the banana peel, which is the main subject.

Figure 15: A POPE object presence question: CM-FDD correctly confirms that a banana is present in the image, in agreement with SFT-1.5B (also "Yes"), while the teacher incorrectly denies its presence, claiming the image only shows a banana peel on a table.