DEPLATENT: VISUAL DEPENDENCY-GUIDED LATENT MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Explicit chain-of-thought (CoT) improves multimodal reasoning but incurs substantial inference cost with long thinking traces. Continuous latent reasoning reduces the number of generated tokens, yet it often underperforms explicit CoT on visually grounded tasks involved with multi-step planning. We first investigate this performance gap through fine-grained error analysis and image-masking interventions, suggesting that latent states may be weakly tied to task-relevant visual evidence. In this paper, we introduce DepLatent, a dependency-guided latent supervision framework that uses region-wise image masking during explicit CoT reasoning to estimate dependencies between reasoning tokens and visual regions. These estimates provide two complementary training signals: dependency-weighted alignment between latent states and relevant visual features, and step-level reconstruction of CoT segments identified by changes in visual dependency. At inference, the multi-modal model reasons using compact continuous latent tokens, avoiding lengthy CoT generation. Across six multimodal reasoning benchmarks and three vision-language backbones with varied sizes, DepLatent consistently improves reasoning accuracy over explicit CoT and latent reasoning baselines while significantly reducing inference latency. Notably, it achieves an average 6.2% improvement over the original Qwen2.5-VL-7B model and speedups over explicit CoT of 2.47$\times$ on the BLINK benchmark.

1 INTRODUCTION

Multi-modal large reasoning models (MLRMs) have demonstrated remarkable capabilities on challenging multi-modal tasks that require long-horizon and multi-step reasoning over visual elements (Deng et al., 2025; Yuan et al., 2025; Yang et al., 2025c). A key driver of this progress is explicit "chain-of-thought" (CoT) training, where MLRMs are typically optimized on long reasoning paths with a next-token prediction loss (Huang et al., 2026; Zhou et al., 2026). These reasoning paths are comprised of descriptive textual steps or interleaved visual tokens, developing the model to decompose complex problems into a sequence of intermediate steps. Despite their advantage, explicit multimodal reasoning approaches often incur substantial inference latency and increased memory consumption (Thawakar et al., 2025; Fan et al., 2026) due to their long thinking process, making deployment on compute- and memory-constrained devices difficult.

To address the efficiency challenges associated with the highly redundant CoT trace, *latent multimodal reasoning* equips MLRMs with compact latent embeddings defined by continuous vectors, serving as intermediate thoughts (Li et al., 2026; Qin et al., 2025; Xu et al., 2026b). The efficacy of these methods relies on supervision to learn a powerful latent representation, which can be broadly categorized into visual-centric and text-centric approaches. The former aligns latent embeddings with those of static visual representations (Li et al., 2026) to enrich latents with visual perception capability, while the latter strengthens the textual reasoning ability of latents by decoding them back into textual sequences with an LLM interpreter (Shen et al., 2025). These methods improve efficiency by reducing token generation, yet their performance remains significantly inferior to the explicit alternative, struggling with complex visually grounded and multi-hop reasoning tasks in benchmarks including EMMA (Hao et al., 2025) and MathVerse (Zhang et al., 2024).

This performance gap raises a fundamental question: *To what extent do latents encode reasoning information comparable to explicit CoT traces?* To gain deeper insights into the representational ca-

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.

pability of latent, we conduct a comprehensive empirical analysis between latent-based strategy and its explicit counterparts, leading to two key observations: (1) Error-type analysis shows that latent-based models have weaker visual perception and logical coherence capabilities than explicit-based models. (2) Latent states exhibit insufficient visual grounding for inter-modal visual-reasoning interaction with only marginal distributional shifts under image masking. *These observations suggest that latents capture insufficient visual dependency modeling compared with explicit CoT traces.*

In this work, we propose DepLatent, a dependency-guided latent supervision framework that adaptively distills the explicit reasoning knowledge identified by the semantic-aware visual masking into a latent reasoning model. Concretely, our method is composed of three components: (i) a reasoning backbone that alternates between a teacher mode, where full CoT traces are leveraged to distill visual-reasoning dependencies and intra-step evolution patterns, and a student mode, where continuous latent thoughts are equipped for efficient inference; (ii) a fine-grained visual alignment module that injects reasoning-critical visual evidence into latent states via the *saliency-magnitude latent*; and (iii) a step-level logic reconstruction module that captures the progressive shift of visual focus across reasoning steps through the *shift-direction latent*. This yields a visually grounded and dynamic supervision signal that teaches the student to think like its own explicit reasoning while preserving the inference-time efficiency of latent reasoning.

Our contributions are summarized as follows:

- We systematically analyze the properties of latents during the reasoning process, i.e., inter-modal visual-reasoning interaction, which primarily results in the degraded performance. Based on our analysis and observation, we identify the representational bottlenecks that hinder the application of latent multi-modal reasoning models, as well as the opportunities that can be leveraged to improve latent reasoning performance and robustness.
- We propose DepLatent, a dependency-guided latent supervision framework that adaptively distills the explicit reasoning knowledge identified by the semantic-aware visual masking into a latent reasoning model, thereby enhancing visual grounding and reasoning consistency within the latent reasoning process.
- We comprehensively evaluate DepLatent across six multimodal reasoning benchmarks and three vision-language backbones. Experimental results show that DepLatent achieves an average 6.2% improvement over the original Qwen2.5-VL-7B model and speedups over explicit CoT of 2.47$\times$ on the BLINK benchmark.

2 Background and Motivation

Explicit Reasoning in MLRMs. Given the multimodal input with the image $I \in \mathcal{I}_{\text{vis}}$ and the textual query $T \in \mathcal{T}_{\text{text}}$, which is first transformed into aligned visual $\mathbf{X}_{\text{vis}} \in \mathbb{R}^{P\times C}$ and textual features $\mathbf{X}_{\text{text}}$, where P and C denote the number of visual tokens and the channel dimension, respectively. The LLM backbone within MLRMs then autoregressively generates the reasoning trace $\mathbf{o} = [o_1, \cdots, o_M]$ consisting of M discrete tokens, before producing the final answer $\boldsymbol{y}$. Overall, the SFT objective for training reasoning models is:

$$\max_{\theta} \mathop{\mathbb{E}}_{(\mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{o}, \boldsymbol{y})} \left[\log p_\theta(\mathbf{o} \mid \mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}) + \log p_\theta(\boldsymbol{y} \mid \mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{o})\right], \tag{1}$$

Latent Reasoning in MLRMs. Instead of strictly confining MLRMs to think along the explicit trace of discrete tokens, latent reasoning models introduce a sequence of learnable continuous embeddings $\mathbf{Z} \in \mathbb{R}^{L\times C}$ to induce intermediate compact reasoning representations, where $L \ll M$. The training objective of latent-based MLRMs broadly follows:

$$\max_{\theta, \mathbf{Z}} \mathop{\mathbb{E}}_{(\mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{o}, \boldsymbol{y})} \left[\log p_\theta(\boldsymbol{y} \mid \mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{Z}) - \lambda \mathcal{R}(\mathbf{H}, \mathbf{o})\right], \tag{2}$$

where $\mathbf{H} = \mathcal{F}_\theta(\mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{Z})$ denotes the final hidden states corresponding to $\mathbf{Z}$ produced by LLM $\mathcal{F}_\theta$. For the supervision module $\mathcal{R}(\cdot, \cdot)$, LVR (Li et al., 2026) adopts the mean squared error between $\mathbf{H}$ and visual embeddings of a predefined fixed region to encode the visual semantics; Heima (Shen et al., 2025) utilizes the decoder to decode $\mathbf{H}$ into textual descriptions. During inference, these continuous *latent tokens* are inserted into the input sequence and contextualized jointly with the

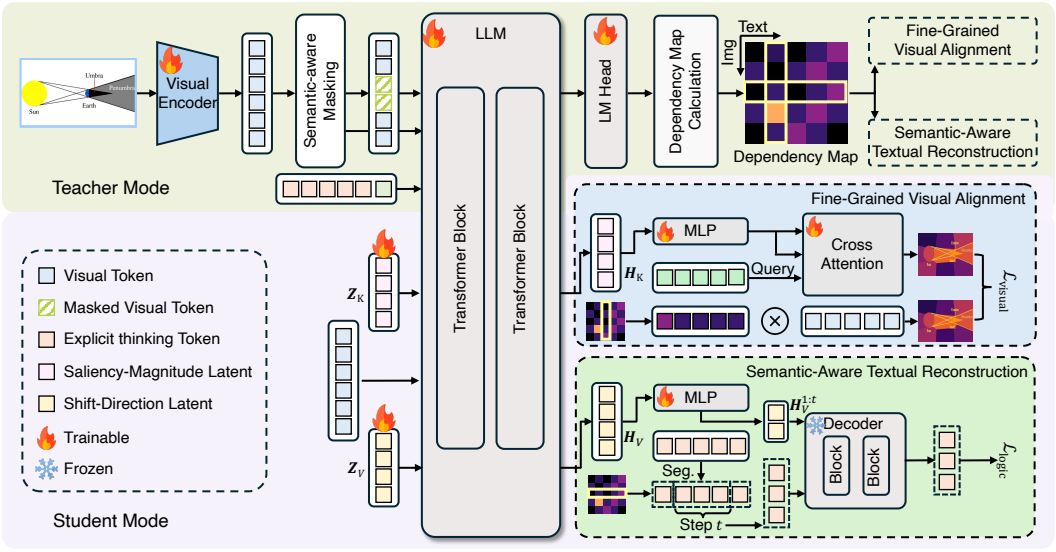


Figure 1: Overview of DepLatent. With a full CoT trace as input in the teacher mode, DepLatent generates the dependency map via the semantic-aware masking operation to learn visual-reasoning dependencies and intra-step evolution. DepLatent then alternates into the student mode to generate continuous compact latent thoughts. During training, the final hidden states of the saliency-magnitude latent $\mathbf{Z}_K$ and shift-direction latent $\mathbf{Z}_V$ are optimized to align with reasoning-critical visual details and to reconstruct the corresponding textual step, respectively, thereby producing a visually grounded and dynamic supervision signal that teaches the latent reasoning model to think like its own explicit reasoning while preserving the inference-time efficiency of latent reasoning.

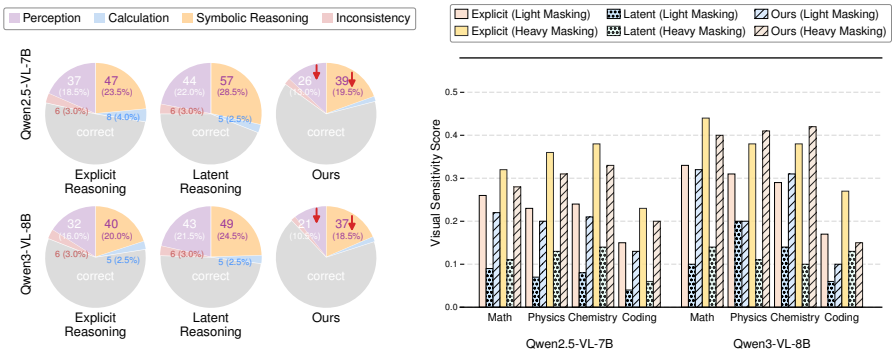
visual and textual tokens. The model then directly generates the final answer without decoding an explicit CoT trace.

Analysis of Explicit vs. Latent Reasoning. To examine whether latent states encode reasoning information comparable to explicit CoT traces, we conduct a comprehensive empirical comparison between latent-based and explicit reasoning models. Rather than only comparing final accuracy, we analyze their reasoning behaviors from three complementary perspectives: outcome-level errors and visual dependency.

Experimental Settings: We follow a typical explicit (Xu et al., 2025) and latent reasoning (Shen et al., 2025) pipeline to train Qwen2.5-VL-7B, Qwen3-VL-8B on Llava-CoT-100K dataset. Note that the experiments are performed under the same backbone model and training data for fairness. We then evaluate them on four benchmarks: EMMA (Hao et al., 2025), MathVista (Lu et al., 2024), MMStar (Chen et al., 2024), BLINK (Fu et al., 2024). These benchmarks cover various perception and reasoning skills across multiple disciplines, including math, biology, physics, chemistry, and coding.

I. Outcome gap: What error patterns dominate in reasoning models? We first randomly sample 200 instances from the above benchmarks and compare the overall error rates of latent reasoning and explicit reasoning. Specifically, latent reasoning yields an overall error rate of 56%, which is substantially higher than the 49% error rate of explicit reasoning on the Qwen2.5-VL-7B backbone. We further analyze and identify four dominant error patterns from these error instances following (Wang et al., 2025), including 1) *perception error*, which stems from inaccurate interpretation of the visual elements; 2) *symbolic reasoning error* due to mistakes in the logical inference process; 3) *calculation error* in performing arithmetic operations; and 4) *inconsistency error* resulting from discrepancies between intermediate reasoning and the final conclusion. As shown in Figure 2 (a), we observe that the majority of errors in the latent-based model are from inaccurate perception (22.0%) and symbolic reasoning (28.5%), such as overlooking key objects and using an invalid logical rule, with both error types appearing substantially more frequently than in the explicit-based model.

II. Visual-dependency gap: How strongly do reasoning states rely on visual evidence? We apply random image masking with ratios of 30% and 60% to introduce light and heavy visual per-



(a) (b)

Figure 2: (a) Error-type breakdown in 200 samples. The results show that inaccurate perception and symbolic reasoning are the main error patterns. (b) The average Jensen-Shannon divergence between the hidden states of the last layer corresponding to the reasoning process before and after masking. It is observed that latent hidden states are not sensitive to the visual information removal, demonstrating insufficient visual grounding throughout the reasoning process.

turbations, and quantify the visual sensitivity of the reasoning process by measuring the average Jensen-Shannon divergence between the last-layer hidden-state distributions before and after masking in Figure 2 (b). This metric captures how much the model's reasoning representations shift when visual evidence is partially removed. It is seen that the latent-based model exhibits substantially lower visual sensitivity than the explicit reasoning model, indicating insufficient visual grounding throughout the reasoning process.

Based on these observations, we argue that the performance degradation of latent reasoning models stems from the lack of incentives to form visually grounded thinking processes. This naturally raises a key question: *can we jointly enhance visual grounding and reasoning consistency within the latent reasoning process?* We present our approach in the next section.

3 METHOD

3.1 OVERVIEW.

In this section, we propose DepLatent, a dual factorized latent supervision framework that successfully distills the rich dependencies and evolution knowledge identified by the semantic-aware visual masking into a latent reasoning model. The overall workflow of DepLatent is illustrated in Figure 1. During training, the input image-question query $[I, T]$ is first transformed into aligned visual $\mathbf{X}_{\text{vis}} \in \mathbb{R}^{P \times C}$ and textual features $\mathbf{X}_{\text{text}}$. We then insert two groups of learned continuous embeddings, $\mathbf{Z}_K$ and $\mathbf{Z}_V$, into the multimodal input sequence. *The saliency-magnitude latent* $\mathbf{Z}_K \in \mathbb{R}^{L_k \times C}$ operates on the fine-grained features space to extract reasoning-relevant visual information. In parallel, *the shift-direction latent* $\mathbf{Z}_V \in \mathbb{R}^{L_v \times C}$ enriches the unique semantic content of each latent by aligning each implicit token with its corresponding explicit reasoning step. In inference, the runtime is essentially that of direct answer generation plus latents forward, which is far shorter than explicit CoT token lengths.

3.2 DEPENDENCY IDENTIFICATION WITH SEMANTIC-AWARE MASKING.

In the teacher mode, we construct the dependency map of the explicit reasoning tokens on the visual regions by performing semantic-aware visual masking. Concretely, given the input image I, we first extract spatial-aware patch tokens $\mathbf{X}_{\text{vis}}$, which is then partitioned into N semantically coherent regions via agglomerative clustering with Ward's linkage, encoding a binary membership matrix $\mathbf{W} \in \{0,1\}^{P \times N}$, where $\mathbf{W}[i, n] = 1$ if and only if vision token i belongs to n-th region. The number of regions N is set adaptively, typically $N \in [16, 48]$. To evaluate the dependency score on vision tokens at n-th region, we set the attention scores from all query positions to it to $-\infty$ before the softmax normalization across all layers and heads within the LLM backbone. This masking-based intervention preserves the original token positions and model state while cleanly removing

the target region's influence on the subsequent generation. Finally, the dependency degree of m-th explicit reasoning token on n-th visual region is measured as the drop in log-probability of the generated token:

$$\mathbf{S}[m,n] = \max(\log \frac{p_\theta\left(o_m \mid o_{<m}, \mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}\right)}{p_\theta\left(o_m \mid o_{<m}, \mathbf{X}_{\text{vis}} \backslash \mathbf{X}_{\text{vis}}^n, \mathbf{X}_{\text{text}}\right)}, 0), \tag{3}$$

where $\mathbf{X}_{\text{vis}} \backslash \mathbf{X}_{\text{vis}}^n$ denotes the masking operator that prevents the model from accessing visual information in the n-th region.

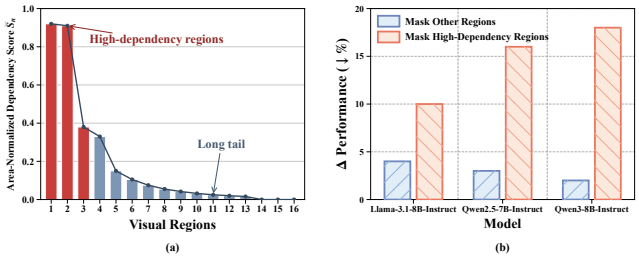


Figure 3: Analysis of visual dependency map using Qwen2.5-VL-7B on EMMA benchmark.

Analysis: Visual regions contribute inequally to the reasoning process. To analyze how visual evidence is used during reasoning, we aggregate and normalize the token-region dependency scores over all explicit reasoning tokens. As shown in Figure 3 (a), we observe a long-tailed distribution: only a small subset of visual regions receives high dependency scores, while most regions contribute marginally. To further verify the functional role of these high-dependency regions, we compare the effect of masking high- and low-dependency regions in Figure 3 (b). Masking high-dependency regions leads to a significant drop in reasoning performance, whereas masking low-dependency regions causes only minor degradation. This indicates that the model relies on a sparse set of critical visual regions during reasoning.

3.3 Saliency-Magnitude Latent with Fine-Grained Visual Alignment.

Instead of aligning latents with all visual regions uniformly, we develop the "*Fine-Grained Visual Localization-Projection-Alignment*" strategy in the student mode to adaptively inject saliency-magnitude latent with reasoning-relevant visual information identified by the masking operator in the projected shared space, thereby avoiding over-regularizing irrelevant or weakly related visual regions.

1 **Saliency-Magnitude Latent Projection.** We extract hidden states corresponding to $\mathbf{Z}_K$: $\mathbf{H}_K = \mathcal{F}_\theta(\mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{Z}_K, \mathbf{Z}_V)[0{:}L_k]$ from the last layer of MLRM $\mathcal{F}_\theta$, and project them through a learnable latent projector $\mathit{Proj}(\cdot)$ to match the latent space with visual representation space. $\mathit{Proj}(\cdot)$ consists of a single linear layer and a cross-attention layer. Specifically, a learnable parameter $\mathbf{Q}$ is used as the query to attend to the mapped latent features, which serve as the keys and values in the cross-attention operation. Therefore, the forward process within the projector can be defined as:

$$\mathbf{Q} \leftarrow \mathit{Proj}\left(\mathbf{Q}, \mathbf{H}_K\right) = \text{CrossAttn}(\mathbf{Q}, \text{MLP}(\mathbf{H}_K)), \tag{4}$$

where $\text{CrossAttn}(\cdot,\cdot)$ denotes the cross attention computations.

2 **Fine-Grained Visual Alignment.** Given the membership matrix $\mathbf{W} \in \mathbb{R}^{P\times N}$, we use the n-th column of $\mathbf{W}$ as a binary mask to select the tokens belonging to region n. Specifically, the region-level representations are obtained as:

$$\bar{\mathbf{Q}}_n = \frac{\sum_{p=1}^{P} \mathbf{W}[p,n]\mathbf{Q}[p,:]}{\sum_{p=1}^{P} \mathbf{W}[p,n]}, \qquad \bar{\mathbf{X}}_n = \frac{\sum_{p=1}^{P} \mathbf{W}[p,n]\mathbf{X}_v[p,:]}{\sum_{p=1}^{P} \mathbf{W}[p,n]}. \tag{5}$$

The dependency-aware visual alignment loss is then computed as a saliency-weighted region-level cosine similarity:

$$\mathcal{L}_{\text{visual}} = \sum_{n=1}^{N} \frac{\mathbf{S}[m,n]}{\sum_{m,n} \mathbf{S}[m,n]} \cos\left(\bar{\mathbf{Q}}_n, \bar{\mathbf{X}}_n\right), \tag{6}$$

where $\cos(\cdot,\cdot)$ denotes the conventional cosine similarity function. By aligning with visual features dynamically, saliency-magnitude learns to encode visual information relevant to reasoning, thereby enabling comprehensive visual reasoning. Note that the alignment is applied only during training, while at inference time the model performs latent mode reasoning without visual alignment supervision.

3.4 Shift-Direction Latent via Step-Level Logic Reconstruction.

During multi-step visual reasoning, different reasoning stages usually rely on visual information in different ways, and the model's dependency on visual regions tends to evolve when the reasoning process moves from one step to another. Therefore, we design the step-level logic reconstruction scheme for the shift-direction latent to ensure the contextual coherence of the reasoning process. We segment the explicit long-thinking trace into individual steps by using period characters as delimiters. We then aggregate token-level visual dependency scores into sentence-level scores, and estimate the dependency-score variation for each sentence boundary based on the difference between neighboring sentence-level scores. We rank all sentence boundaries by their variation magnitude and select the top-ranked boundaries as segmentation points. Therefore, the reasoning process is divided at these selected boundaries to obtain the desired number of segments $\mathbf{O} = \{\mathbf{o}^1, \cdots, \mathbf{o}^{L_v}\}$. For t-th step, the LLM decoder p_φ takes the first t implicit hidden states $\mathbf{H}_V^{0:t} = \mathcal{F}_\theta(\mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{Z}_K, \mathbf{Z}_V)[L_k{:}L_k + t]$ as a conditioning signal and generates the corresponding reasoning step $\mathbf{o}^t$. Since $\mathbf{H}_V^{0:t}$ is not a vocabulary token, the loss is computed only over the textual step tokens:

$$\mathcal{L}_{\text{logic}} = -\sum_{t=1}^{L_k} \log p_\varphi(\mathbf{o}^t \mid \mathbf{H}_V^{0:t}). \tag{7}$$

This loss grounds each latent $\mathbf{Z}_V^k$ to a specific reasoning step, ensuring that the latent sequence carries fine-grained semantics. The decoder is used exclusively for this supervision during training and is discarded at inference.

3.5 Overall Objective.

Overall, the final training loss $\mathcal{L}$ involves three components to empower the saliency-magnitude latent with fine-grained visual grounding, develop the shift-direction latent via step-level semantic reconstruction, and supervise the generation of the final answer $\boldsymbol{y}$, respectively:

$$\mathcal{L} = \lambda_{\text{logic}}\mathcal{L}_{\text{logic}} + \lambda_{\text{visual}}\mathcal{L}_{\text{visual}} + \lambda_{\text{ans}}\mathcal{L}_{\text{ans}}, \tag{8}$$

where $\mathcal{L}_{\text{ans}} = -\log p_\theta(\boldsymbol{y} \mid \mathbf{X}_{\text{vis}}, \mathbf{X}_{\text{text}}, \mathbf{Z}_K, \mathbf{Z}_V)$.

4 Experiments

4.1 Experimental Settings

Implementation Details. We adopt the LLaMA and Qwen series models across different scales as the base MLLM, including Llama-3.2-11B-Vision-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B-Instruct (Bai et al., 2025a). Following Heima (Shen et al., 2025), we finetune the base MLLM on LLava-CoT-100K (Xu et al., 2025) with a LoRA setup of rank 16, alpha value 32, and dropout 0.01 for all the experiments. The Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen3-8B-Instruct (Yang et al., 2025a) is employed as the decoder, respectively. For the cross attention module within the saliency-magnitude latent projector, we set the number of head as 8. The training is conducted on the servers equipped with 8 NVIDIA RTX PRO 6000 GPUs, 2 AMD EPYC 9254 CPUs (96 Cores), and 1.5 TB of RAM.

Baselines. We compare our framework with three types of baselines: (1) Zero-shot VLMs including Qwen3-VL-8B (Bai et al., 2025a), Qwen2.5-VL-7B (Bai et al., 2025b); (2) Explicit multi-modal reasoning works: LLava-CoT (Xu et al., 2025), Qwen3-VL-8B + vanilla SFT, Qwen2.5-VL-7B + vanilla SFT; (3) Latent multi-modal reasoning approaches: CoVT (Qin et al., 2025), LVR (Li et al., 2026), Heima (Shen et al., 2025).

Table 1: The results of our model and baselines on multi-modal reasoning benchmarks.

Method	Training Set Size	Multi-discipline		Math-Related		Vision-centric				Overall
		EMMA	MMMU(val)	MathVista	MathVerse	MMStar	JigSaw	Rel. Refl.	Rel. Depth	AVG
Llama-3.2-11B-Vision-Instruct										
LLava-CoT	100K	19.5	22.4	50.9	28.3	54.0	51.0	-	-	37.6
Heima	100K	20.4	20.6	43.6	25.5	50.9	49.5	-	-	35.0
Ours	100K	**30.7**	**25.7**	**53.1**	**30.6**	**55.2**	**57.2**	-	-	**42.1**
Qwen3-VL										
Qwen3-VL-8B	-	24.6	69.2	74.5	61.4	68.2	71.0	42.3	85.9	62.1
+ vanilla SFT	100K	26.2	70.3	**79.3**	63.4	64.0	75.6	41.5	82.6	62.9
Heima	100K	24.1	62.1	71.6	56.6	62.5	70.3	44.2	82.5	59.2
Ours	100K	**34.8**	**70.4**	77.2	**64.6**	**70.3**	**76.2**	**49.9**	**89.6**	**66.6**
Qwen2.5-VL										
Qwen2.5-VL-7B	-	21.5	57.1	68.2	45.7	63.3	52.0	38.8	79.8	53.3
+ vanilla SFT	100K	23.1	**59.4**	67.8	**48.3**	63.1	60.4	40.5	80.9	55.4
CoVT	889K	13.5	51.3	67.3	39.4	**69.2**	60.3	43.7	82.7	53.4
LVR	477K	11.4	54.0	42.6	38.7	64.4	52.7	41.0	81.4	48.2
Heima	100K	22.6	49.7	70.4	40.5	62.1	59.2	41.5	80.9	53.6
Ours	100K	**34.3**	57.3	**72.7**	46.3	67.5	**68.7**	**45.0**	**84.1**	**59.5**

Benchmarks. To thoroughly evaluate the effectiveness of DepLatent, we conduct experiments on a diverse set of challenging multimodal benchmarks, including: (1) Multi-discipline benchmark: EMMA (Hao et al., 2025), MMMU (Yue et al., 2024), which cover common subjects such as physics, coding, mathematics, chemistry, art and design, business, and science; (2) Math-Related benchmark that covers various mathematical fields to evaluate MLRMs' math reasoning ability: MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024); (3) Vision-centric benchmark to assess the general visual reasoning ability of our model: MMStar (Chen et al., 2024), BLINK (Fu et al., 2024). All evaluation metrics are based on exact match against the ground truth answer.

4.2 Main Results

DepLatent outperforms baselines with consistently improved reasoning accuracy. We present the main experimental results in Table 1. Compared to the original Qwen2.5-VL-7B base models, our approach achieves a substantial improvement, with an average increase of approximately 6.2% in accuracy across all tasks. Furthermore, it surpasses both Vanilla SFT and Heima fine-tuned with the same long chain-of-thought data on most benchmarks. One possible reason is that Vanilla SFT optimizes all reasoning tokens with a uniform next-token prediction objective, causing the training signal to be dominated by long CoT traces rather than the final answer. As a result, the model may overfit redundant or noisy reasoning patterns, which can dilute answer-level supervision and even hurt final accuracy. Our method is effective across all model scales, from the smallest 7B model to the largest 11B model, highlighting its strong generalization ability to different model capacities and diverse benchmarks.

Substantially reduced perception and symbolic reasoning errors. We further conduct a comprehensive qualitative study on the error distribution, following the setup detailed in Section 2. Figure 2 shows a before-and-after comparison of errors. We observe a significant reduction in perception and symbolic reasoning errors, demonstrating the effectiveness of addressing the perception and symbolic reasoning bottleneck in latent-based multi-modal reasoning.

Efficiency with significant speedup. We report the latency of DepLatent during inference on the BLINK and MathVerse benchmarks, as shown in Table 2. LLava-CoT and Qwen2.5-VL-7B + vanilla SFT directly finetunes the base VLM model on long and redundant CoT data, thereby suffering from slow inference. In comparison, our DepLatent leverages the latent with compact thinking information to speed up the inference process without compromising the performance.

Table 2: Latency (s) on BLINK and MathVerse benchmarks.

Method	BLINK	MathVerse
Llama-3.2-11B-vision	5.9	13.05
LLava-CoT	12.5 (1.00 ×)	16.8 (1.00 ×)
Ours	**2.1** (5.95 ×)	**0.93** (18.06 ×)
Qwen2.5-VL-7B	1.4	11.1
Qwen2.5-VL-7B + vanilla SFT	1.8 (1.00 ×)	15.5 (1.00 ×)
Ours	**0.73** (2.47 ×)	**0.4** (38.75 ×)

Table 3: The ablation of the latent components and the alignment strategy with Qwen2.5-VL-7B backbone.

Method	EMMA	MathVista	Rel. Refl.	Rel. Depth.	Avg.
Qwen2.5-VL-7B	21.5	68.2	38.8	79.8	52.1
Saliency-Magnitude Latent Only	32.8	71.5	42.6	82.0	57.2
Shift-Direction Latent Only	30.6	70.3	39.1	78.2	54.6
Heima	22.6	70.4	41.5	80.9	53.8
Ours	**34.3**	**72.7**	**45.0**	**84.1**	**58.5**

4.3 Ablation Study

Ablation on the Latent Components. To evaluate the effectiveness of DepLatent in integrating saliency-magnitued latent and shift-direction latent, we compare DepLatent with saliency-magnitude latent-only and shift-direction latent-only strategies on EMMA, MathVista, and BLINK benchmarks in terms of end-to-end reasoning quality. We report results in Table 3, and observe that DepLatent achieves the best reasoning accuracy among other alternatives, confirming the effectiveness of our latent fusion strategy.

Ablation on the Alignment Strategy. We study the importance of our proposed fine-grained visual alignment vs. other alignment variants, including the uniform alignment strategy and the random alignment strategy. As shown in Figure 4 (a), our approach significantly outperforms both variants across all benchmarks, indicating the effectiveness of our reasoning-relevant visual information injection. Moreover, the uniform alignment alternative achieves the worst performance on the MathVista benchmark. This is because the math-related images often contain substantial redundant or reasoning-irrelevant visual regions, such as blank backgrounds. Enforcing uniform alignment over all visual tokens may dilute the model's attention to reasoning-relevant visual cues, making it harder to focus on regions that are critical for answering the question.

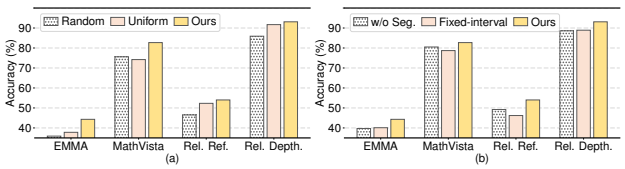


Figure 4: (a) Ablation study of different visual alignment designs. (b) Comparison of different CoT trace segmentation strategies.

Ablation on the Segmentation Strategy. We further evaluate the effectiveness of the proposed segmentation strategy by comparing it with two alternatives: no segmentation and fixed-interval segmentation with an interval of 50 tokens. As shown in Table 4 (b), our segmentation improves over no segmentation by 3.1% on EMMA, 3.5% on Rel. Refl., and 4.5% on Rel. Depth. These improvements highlight that the dependency-aware segmentation effectively directs the model toward more promising reasoning process by providing more coherent and informative latent components for alignment.

Table 4: Ablation study of decoder selection strategy.

Method	EMMA	MathVista	Rel. Refl.	Rel. Depth	Avg.
Heima	22.6	70.4	41.5	80.9	53.8
3B Decoder	30.5	70.9	43.6	81.9	56.7
7B Decoder	**34.3**	**72.7**	**45.0**	**84.1**	**59.0**
14B Decoder	34.0	71.5	43.8	83.1	58.1

Ablation on the Decoder. We replace the decoder of Qwen2.5-7B with smaller and larger ones from the same family and evaluate them on EMMA, MathVista, and BLINK benchmarks. These results are reported in Table 4. Compared with Heima, all decoder variants bring consistent improvements, indicating the effectiveness of incorporating a stronger decoder for latent reasoning supervision. Among them, the 7B decoder achieves the best average performance, while the 3B decoder performs worse. This may be attributed to two factors: first, the 3B decoder has a different vocabulary size from the 7B backbone, which may introduce vocabulary-level mismatch during reconstruction; second, its smaller parameter scale may limit its expressive capacity for modeling complex reasoning steps. In contrast, the 14B decoder achieves performance comparable to the 7B decoder, suggesting that further increasing the decoder size brings marginal gains once sufficient modeling capacity is reached.

Table 5: Ablation study of latent number on overall performance. The green color indicates our default setting.

Num L_k	EMMA	MathVista	Rel.Refl	Rel.Depth	Num L_v	EMMA	MathVista	Rel.Refl	Rel.Depth
4	29.9	69.3	44.8	83.8	1	29.7	70.5	39.2	78.6
6	33.7	69.8	42.5	83.1	2	30.8	71.3	40.5	83.1
8	**34.3**	72.7	45.0	**84.1**	3	33.6	72.7	44.7	80.6
10	33.2	72.5	**45.1**	84.0	4	**34.3**	72.7	**45.0**	**84.1**
12	34.3	**72.8**	44.7	84.1	5	33.9	**72.8**	43.6	84.1

Ablation on the Number of Latent Tokens. The length of saliency-magnitude latent and shift-direction latent are controlled by the hyperparameters L_k and L_v. In order to determine the appropriate settings, we explore the influence of L_k and L_v on performance on EMMA, MathVista, and BLINK benchmarks. As demonstrated in Table 5, for L_k, increasing it from 4 to 8 improves most metrics, while further increasing it to 10 or 12 brings only marginal fluctuations and no consistent gains. Similarly, for L_v, the performance improves as it increases from 1 to 4, but slightly drops when $L_v = 5$. Therefore, we set $L_k = 8$ and $L_v = 4$ as the default configuration, which achieves the best overall trade-off across all metrics.

5 Related Work

Explicit Multimodal Reasoning. Explicit multimodal reasoning generates intermediate reasoning steps in a discrete, interpretable form before producing the final answer. Early methods primarily rely on textual chain-of-thought (CoT), but text-only rationales often provide insufficient visual grounding, leading to cross-modal misalignment and hallucination. To address this issue, subsequent approaches convert visual content into textual descriptions using captioning models, external tools, or specialized visual experts. For example, Hu et al. extract image semantics as captions and concatenate them with the input prompt to support downstream reasoning (Hu et al., 2024). More recent methods instead adopt interleaved vision-text rationales, allowing models to manipulate visual inputs through drawing, marking, zooming, cropping, or code-based editing during reasoning (Yang et al., 2025b). Other approaches generate auxiliary images with visual annotations and combine them with textual steps. Xu et al. perform reasoning entirely through dynamically generated images, demonstrating substantial gains on visual navigation tasks (Xu et al., 2026a).

Latent Reasoning. Unlike explicit reasoning over discrete tokens, latent reasoning performs intermediate computation directly in the model's continuous hidden space. Hao et al. pioneer this paradigm by recurrently feeding the previous hidden state as the input embedding for the next reasoning step, thereby avoiding the generation of lengthy textual rationales and reducing inference latency (Hao et al., 2024). However, without direct supervision, intermediate latent states may become homogeneous or fail to encode distinct reasoning steps. Subsequent work therefore improves latent representations through more flexible architectures and targeted supervision. Recent studies extend latent reasoning to multimodal large language models, where the central challenge is to preserve and integrate task-relevant visual information in the latent space (Yue et al., 2026). Li et al. address this challenge through a structured cognitive hierarchy that emphasizes fine-grained image details, although their method relies on annotation-intensive multimodal reasoning data (Li et al., 2026).

6 Conclusion

In this paper, we introduce DepLatent, a dependency-guided latent supervision framework that uses region-wise image masking during explicit CoT reasoning to estimate dependencies between reasoning tokens and visual regions. Across six multimodal reasoning benchmarks and three vision-language backbones with varied sizes, DepLatent consistently improves reasoning accuracy over explicit CoT and latent reasoning baselines while significantly reducing inference latency. It achieves an average 6.2% improvement over the original Qwen2.5-VL-7B model and speedups over explicit CoT of 2.47$\times$ on the BLINK benchmark.

AI USE STATEMENT

The authors used OpenAI Codex to improve the clarity, grammar, and readability of the manuscript and to assist in writing code for portions of the data analysis. All AI-generated code was reviewed, tested, and executed by the authors, and the resulting analyses were independently verified. The authors take full responsibility for the final content, claims, and conclusions of this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 123817–123846, 2025.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Wang. Grit: Teaching mllms to think with images. *Advances in Neural Information Processing Systems*, 38:116522–116543, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced multimodal reasoning benchmark. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=v26vwjxOEz.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UZIjskfbfU.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=j84WR5ORsC.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Xuan Shen, Yizhou Wang, Yufa Zhou, Xiangxi Shi, Pu Zhao, Yanzhi Wang, and Jiuxiang Gu. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*, 2025.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24290–24315, 2025.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. In *International Conference on Learning Representations*, volume 2026, pp. 36706–36739, 2026a.

Zhongxing Xu, Zhonghua Wang, Zhe Qian, Dachuan Shi, Feilong Tang, Ming Hu, Shiyan Su, Xiaocheng Zou, Wei Feng, Dwarikanath Mahapatra, et al. Thinking in uncertainty: Mitigating hallucinations in mlrms with latent entropy-aware decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11064–11075, 2026b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025b.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2376–2385, 2025c.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*, 2025.

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*, pp. 9556–9567, 2024.

Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. Hybrid latent reasoning via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:5501–5530, 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=28qUA2bSe5.

A APPENDIX

A.1 THE CLUSTERING QUALITY.

Implementation Details To define interpretable visual regions for attribution, we perform spatially-constrained agglomerative clustering on the extracted patch features.

We use cosine distance between feature vectors to measure semantic dissimilarity, and Ward's linkage is used to minimize the variance within clusters. Instead of a fixed number of clusters, we dynamically determine the number of regions N for each image based on a distance threshold.

Comparison with K-means Clustering. We also compare agglomerative clustering with K-means clustering using fixed N values. Table 6 shows that agglomerative clustering with adaptive N outperforms K-means across all fixed N settings, particularly at lower N values where semantic boundaries become critical.

Table 6: Comparison of fixed-N K-means and adaptive agglomerative clustering on EMMA using Qwen3-VL-8B.

Method	Math	Physics	Chemistry	Coding
K-means ($N = 16$)	31.1	40.4	29.8	31.0
K-means ($N = 32$)	33.3	42.9	31.7	33.0
K-means ($N = 48$)	34.0	43.6	32.4	33.7
K-means ($N = 64$)	33.4	42.3	31.9	33.2
Ours (adaptive)	**35.3**	**44.9**	**33.2**	**34.6**

A.2 TRAINING COMPUTATION COSTS ANALYSIS.

As shown in Table 7, DepLatent introduces a limited increase in training time across different backbone models. Compared with Heima, DepLatent requires an additional 2 hours for Qwen2.5-VL-7B, 4 hours for Qwen3-VL-8B-Instruct, and 16 hours for Llama-3.2-11B-Vision. These results indicate that the additional computational cost varies with model scale but remains manageable. Given the consistent performance improvements achieved through semantic-aware dependency modeling, this additional training cost represents a favorable trade-off.

Table 7: Training cost analysis.

Method	Backbone	Time (h)
Heima	Llama-3.2-11B-Vision	48
DepLatent		64
Heima	Qwen2.5-VL-7B	24
DepLatent		26
Heima	Qwen3-VL-8B-Instruct	27
DepLatent		31