Visual Enhanced Depth Scaling for Multimodal Latent Reasoning

Yudong Han[1,2,3]*, **Yong Wang**[2]*, **Zaiquan Yang**[4], **Zhen Qu**[5], **Liyuan Pan**[1]†, **Xiangxiang Chu**[2]
[1]Computer Science, Beijing Institute of Technology, [2]DreamX Team,
[3]State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China,
[4]City University of Hong Kong,
[5]Institute of Automation, Chinese Academy of Sciences,
Code: https://github.com/Simon98-AI/Vedas

Abstract

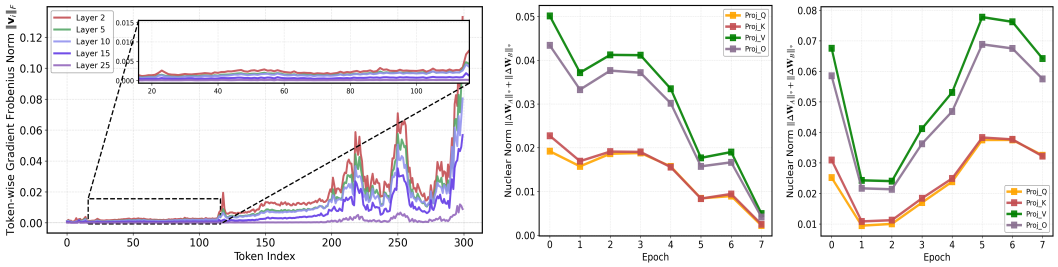
Multimodal latent reasoning has emerged as a promising paradigm that replaces explicit Chain-of-Thought (CoT) decoding with implicit feature propagation, simultaneously enhancing representation informativeness and reducing inference latency. By analyzing token-level gradient dynamics during latent training, we reveal two critical observations: (1) visual tokens exhibit significantly smaller gradient norms than their textual counterparts due to inherent language bias, resulting in systematic visual under-optimization; and (2) semantically simple tokens converge rapidly, whereas complex tokens exhibit persistent gradient instability constrained by fixed architectural depths. To address these limitations, we propose a visual replay module and routing depth scaling to collaboratively enhance visual perception and refine complicated latents for deeper contextual reasoning. The former module leverages causal self-attention to estimate token saliency, reinforcing fine-grained grounding through spatially-coherent constraints. Complementarily, the latter mechanism adaptively allocates additional reasoning steps to complex tokens, enabling deeper contextual refinement. Guided by a curriculum strategy that progressively internalizes explicit CoT into compact latent representations, our framework achieves state-of-the-art performance across diverse benchmarks while delivering substantial inference speedups over explicit CoT baselines.

1 Introduction

Over the past few years, large language models (LLMs) have achieved remarkable progress in complex reasoning [11, 19, 54, 89, 111], demonstrating substantial capabilities across diverse applications such as mathematics [46, 72, 88], code generation [79, 90], and AI for science [102, 104, 105], propelled by scaling laws in data volume [1, 66] and model capacity [29, 60]. Advanced techniques such as Chain-of-Thought (CoT) prompting [57, 108] and reinforcement learning (RL) [9, 11, 14, 97] for trajectory optimization have proven highly effective in text-only domains. Extending these capabilities to the multimodal realm has thus become a pivotal research direction. Current approaches primarily follow three paradigms. First, *Text-based Reasoning* [30, 49, 108] generates explicit multi-step textual chains before producing an answer. However, these methods typically rely on static visual inputs, and recent studies [92, 98] indicate that visual grounding deteriorates significantly over extended reasoning chains. Second, *Tool-augmented Reasoning* manipulates visual inputs through external operations (e.g., zooming or region enhancement) and injects intermediate visual hints into the reasoning trace. While powerful, these approaches are prone to redundant or invalid tool invocations,

* Equal contribution. Work done when Yudong's internship at DreamX Team.
† Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Token Frobenius Norm in Different Epochs (b) Easy Samples (c) Hard Samples

Figure 1: Panel (a) depicts the token-wise Frobenius norm of gradients across different layers throughout the entire training process, revealing a clear gradient imbalance between visual tokens and text tokens/latent token. Panels (b) and (c) further reveal distinct evolution patterns of the Nuclear norm in the QKVO projection matrix across layers and training epochs: gradients for easy samples decay rapidly and converge smoothly, whereas hard samples maintain elevated gradient norms with persistent oscillations. Here, $\mathbf{W}_A$ and $\mathbf{W}_B$ denote the low-rank matrix in LoRA [22].

introducing noise that degrades performance and substantially increases inference latency. Recently, emerging researches position *Latent Reasoning* as a viable direction for optimizing multimodal reasoning. Unlike traditional methods that rely on explicit textual chains, latent reasoning encodes intermediate reasoning steps into compact continuous vectors. This paradigm offers compelling advantages, including higher inference efficiency, reduced annotation overhead, and the ability to learn dense, high-fidelity multimodal representations [36, 37, 62–64, 108].

To gain deeper insights into the optimization dynamics of latent reasoning, we conducted a systematic analysis of gradient flows and parametric evolution during training. This investigation reveals two meaningful observations: *(1) Visual-Text Optimization Disparity:* Recent studies [81, 91, 101] have highlighted the phenomenon of visual attention attenuation in MLLMs as the explicit Chain-of-Thought (CoT) reasoning chain extends. We discover that a similar degradation existing in latent reasoning. As depicted in Fig. 1a, the gradient updates for visual tokens (regions marked by dashed lines) are remarkably small, while the overall optimization is predominantly governed by gradients from text tokens and latent token. This suggests that the model relies primarily on textual information and latent representations to generate final answers during training, revealing a pronounced gradient imbalance phenomenon. *(2) Fixed-Depth Optimization Dilemma:* Beyond modality imbalance [83, 103], we observe an architecture bottleneck in how models handle samples with different complexity. Partitioning the training data into *easy* (consistently correct) and *hard* (persistently incorrect) subsets based on early-stage validation accuracy, we tracked the gradient nuclear norms across the QKV and output projection matrices O (Fig. 1b and Fig. 1c). The trends reveal a stark divergence: easy samples exhibit smooth gradient decay, indicating stable convergence into favorable loss basins. In contrast, hard samples maintain persistently high gradient volatility even in late training epochs. This phenomenon underscores the necessity of iterative refinement for complex reasoning, which is crucial for parsing compositional patterns in complex contexts while mitigating inherent visual ambiguities in images. Fixed-depth architectures, however, lack the flexibility to adapt to varying token complexities, thereby trapping hard examples in oscillatory optimization trends.

Motivated by these observations, we propose a unified framework (Vedas) that strengthens fine-grained visual engagement and achieves token-wise depth scaling strategy to enable more precise and comprehensive contextual reasoning. Firstly, to mitigate visual optimization instability, we design a visual replay module, which dynamically replays the focused visual clues interleaved with thinking latents. This mechanism iteratively exposes and propagates key visual context across reasoning steps, fostering step-wise alignment with the target answer. Complementing this, we apply self-distillation supervision to enforce spatial coherence and preserve fine-grained visual details within the visual latents. Secondly, to address the fixed-depth optimization bottleneck, we propose a per-layer token router that dynamically allocates additional reasoning steps based on token complexity or information density. This design enables high-difficulty tokens to engage in prolonged contextual reasoning by reusing layer-wise knowledge, facilitating iterative representation refinement and adaptive prioritization of salient contextual cues. Finally, in contrast to methods [21, 58] that rely on knowledge distillation for direct latent supervision, we employ a curriculum learning strategy that

progressively introduces latent tokens into the training pipeline. Extensive experiments show that our method can be effortlessly combined with various widely-used MLLM backbones to further enhance reasoning performance while maintain the satisfactory inference latency. The main contributions can be summarized as follows:

- We systematically analyze token-level gradient dynamics during latent reasoning training, revealing two critical optimization bottlenecks: visual-text optimization disparity and fixed-depth optimization dilemma.
- We present a unified curriculum-driven framework that progressively constructs interleaved latent representations. By integrating spatially-coherent visual constraints for fine-grained grounding and complexity-aware depth scaling, our approach enables robust and precise contextual reasoning.
- Extensive experiments across twelve widely-used multimodal reasoning benchmarks demonstrate that our method achieves state-of-the-art performance while maintaining high inference efficiency.

2 Related Work

Explicit Multimodal Reasoning. Multimodal reasoning enables model to reason over information from different modalities to solve complex tasks. There are many prior works [17–19, 23, 38, 42, 47, 49, 49, 50, 93, 94, 99, 100, 110] focusing extensively on enhancing reasoning capabilities. Earlier work rely on the CoT prompting to perform explicit thinking steps in the text space before generating the final answer. However, this paradigm generally lack sufficient visual grounding capability and leads to unsatisfactory misalignment and hallucination [4, 25, 68]. To address these limitations, recent studies [23, 49, 110] have explored converting visual information into textual formats prior to reasoning, leveraging external tools or specialized visual experts to generate descriptive representations that guide LLMs. For instance, Hu et al. [23] pioneered the integration of visual captions, extracting semantic content as text and concatenating it with input prompts to bolster reasoning capabilities. To further enhance fine-grained reasoning, subsequent works have focused on regional understanding, aiming to improve textual expressiveness by describing specific image regions. Others [47, 49, 50] identified entities and their relationships within images, facilitating fine-grained reasoning through explicit modeling of inter-entity connections.

A line of concurrent works advocates using vision-text interleaved format during the rationale generation and reasoning process. The model draws auxiliary lines or marks based on original image to record thinking path, zoom or crop regions, or perform code editing, etc. Building on these paradigms, Zhang et al. [108] first proposed decoupling rationale generation from answer generation in the Vision-Text Reasoning field. Subsequently, Shao et al. [56] annotated key regions of the original image in intermediate steps, training models to focus on image regions relevant to the answer. While some works [15, 107] further extract key image regions progressively during reasoning, combining visual information with textual reasoning to generate the final answer. Moreover, new methods [24, 41] emulated human thought by sketching images during reasoning, focusing on core concepts, structures, and relationships while ignoring redundant details. Other works [8, 35, 67] generated new images with auxiliary markers during reasoning, combining them with text to improve reasoning in complex scenarios. To fully shift reasoning from the linguistic domain to the visual modality, Xu et al. [87] proposes to reasoning exclusively with dynamic generated images, demonstrating substantial performance gains in visual navigation tasks.

More recently, Vision-R1 [26] and VL-Rethinker [70] leveraged Group Relative Policy Optimization [11] (GRPO) to refine reasoning trajectories through rollout-based sampling and reward scoring. Complementing these policy-driven approaches, concurrent works further enhance reasoning capabilities via novel cognitive paradigms, including self-critiquing cycles [10, 20, 55], iterative rethinking [92], and on-policy distillation [95, 109].

Latent Reasoning. Different from explicit reasoning in the discrete token space, latent reasoning refers to internal computation performed in a hidden space before answer generation. Hao et al. [21] pioneers continuous latent space reasoning by feeding the last hidden states as input embeddings for the next step without generating intermediate discrete tokens, substantially reducing reasoning latency. However, subsequent studies have indicated that such paradigm may suffer from feature homogenization without explicit supervision on intermediate latent states. To address this limitation, a

series of works attempt to enhance the quality of intermediate representations using diverse strategies. Cheng and Van Durme [7] introduced variable-length contemplation tokens for latent reasoning, mitigating quality degradation caused by fixed-length constraints. Similarly, Shen et al. [59] leveraged distillation tactic to align student and teacher hidden activations along with explicit supervision, thereby constraining latent reasoning paths. Beyond these efforts, Wei et al. [85] adopted step-level supervision to further stabilize the reasoning space.

Recently, latent reasoning has been extended to Multimodal Large Language Models (MLLMs). Distinct from text-only LLMs, MLLMs necessitate the effective integration of visual features within the latent reasoning space. Several efforts [31, 40, 43, 53, 80, 96, 99] have been dedicated to injecting visual cues into the latent space to facilitate visual-grounded reasoning. For instance, Wang et al. [80] emphasize image details by constructing a structured cognitive hierarchy, albeit relying on annotation-intensive multimodal reasoning data. Similarly, Liu et al. [40] progressively select visual patches to inject into latent thinking tokens via confidence-guided policy gradient optimization. In this work, we systematically analyze gradient dynamics during latent reasoning training and reveal two critical bottlenecks towards token-wise optimization behavior.

3 Method

3.1 Preliminary: Implicit and Explicit Decoding

Given a multimodal input comprising a question $\mathcal{Q}$ and an image $\mathcal{V}$, we tokenize them into a text embedding sequence $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N_q}$ and visual features $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N_v}$ using a word embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{W}| \times d}$ and a pretrained visual encoder, respectively. We concatenate $\mathbf{X} = [\mathbf{Q} \| \mathbf{V}]$ and use an autoregressive MLLM $\mathcal{F}_\theta$ to obtain the prefilling hidden states $\mathbf{H}^{(0)} \in \mathbb{R}^{P \times d}$, where $P = N_q + N_v$. During decoding, we fix T_r implicit reasoning steps and T_a explicit answer tokens. The generation process has two stages.

Implicit Reasoning Phase. At step t, the model produces a continuous latent vector $\mathbf{z}^{(t)}$ conditioned on $\mathbf{X}$ and previous latent states:

$$\mathbf{z}^{(t)} = \mathcal{T}(\mathcal{F}_\theta(\mathbf{X} \,\|\, \mathbf{Z}_{<t})), \quad t = 1, \ldots, T_r, \tag{1}$$

where $\mathbf{Z}_{<t} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(t-1)}\}$ and $\mathcal{T}(\cdot)$ extracts the hidden state of the last token. The latent sequence $\mathbf{Z}_{1:T_r}$ is then concatenated to the context for answer decoding.

Explicit Answer Decoding. The answer tokens $\mathbf{a}_{1:T_a}$ are generated autoregressively. For t-th answer token a_t:

$$p_\theta(a_t \mid \mathbf{X}, \mathbf{Z}_{1:T_r}, a_{<t}) = \mathrm{Softmax}\big(\mathbf{W}_o \mathbf{h}_t^{\mathrm{dec}}\big), \tag{2}$$

where $\mathbf{W}_o \in \mathbb{R}^{|\mathcal{W}| \times d}$ is the output projection matrix and $\mathbf{h}_t^{\mathrm{dec}}$ is the decoder hidden state at step t. The joint likelihood factorizes as

$$p_\theta(\mathbf{a}_{1:T_a} \mid \mathbf{X}, \mathbf{Z}_{1:T_r}) = \prod_{t=1}^{T_a} p_\theta(a_t \mid \mathbf{X}, \mathbf{Z}_{1:T_r}, a_{<t}). \tag{3}$$

At each decoding step, the decoder input is $\mathbf{U}_t = [\mathbf{X} \| \mathbf{Z}_{1:T_r} \| a_{<t}]$, with $\|$ denoting concatenation along the token dimension.

3.2 Spatially-Coherent Finer Visual Replay

As mentioned earlier, we empirically reveal the gradient disparities between visual and textual tokens throughout the learning dynamics. Specifically, visual tokens consistently exhibit substantially smaller gradient Frobenius norms compared to textual tokens and latent token, indicating that visual representations remain under-optimized despite their critical role in multimodal reasoning, which aligns with the conclusions drawn in prior studies [52, 84]. Motivated by these insights, we introduce the visual replay module to reinforce the engagement of visual cues via salient region detection, while enhancing fine-grained spatially-coherent perception capabilities via self-distillation supervision at each reasoning step.

Attention-Guided Region Focus. Several works [3, 20, 39, 106] have demonstrated that LLMs exhibit fundamental visual grounding capabilities. To identify visually salient regions, we aggregate

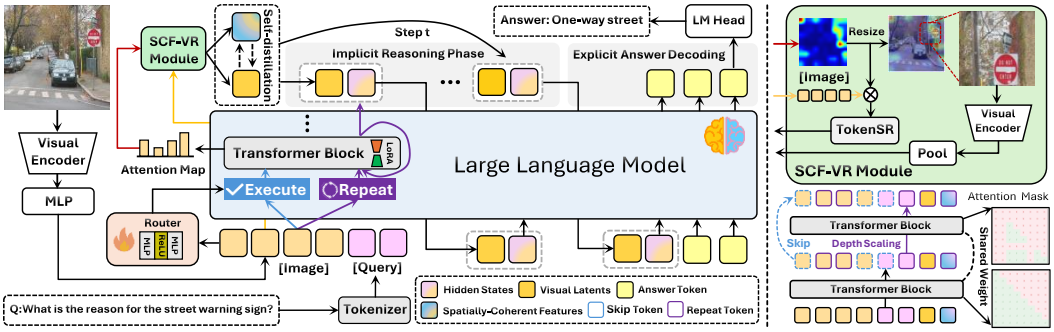


Figure 2: **Left Panel:** Schematic illustration of our framework. Input images are encoded into visual tokens by a pretrained visual encoder and projected into a text-centric semantic space aligned with the LLM, while questions are tokenized by the text tokenizer. Our method enhances a standard latent MLLM with two synergistic components during the implicit reasoning phase: Spatially-Coherent Finer Visual Replay (SCF-VR) and Routing Depth Scaling (RDS). The overall objective is jointly optimized via standard cross-entropy loss and self-distillation loss. **Right Panel:** Detailed architecture of the SCF-VR module and token-wise depth scaling mechanism. **TokenSR** denotes the token super-resolution module, and **Pool** is the average pooling.

attention weights across all transformer layers and attention heads to obtain a consolidated spatial focus map. Specifically, given the l-th layer and the h-th attention head, we compute the mean attention map $\bar{\mathbf{A}}^{(t)}$ at reasoning step t:

$$\bar{\mathbf{A}}^{(t)} = \frac{1}{L \cdot H} \sum_{l=1}^{L} \sum_{h=1}^{H} \mathbf{A}^{(l,h,t)}, \tag{4}$$

where $\mathbf{A}^{(l,h,t)} \in \mathbb{R}^{P^{(t)} \times P^{(t)}}$ denotes the attention matrix for layer l and head h at iteration t ($1 \leq t \leq T_r$), with $P^{(t)}$ representing the number of input tokens at iteration t. Here, L and H represent the total number of layers and heads, respectively. To obtain token-level attention scores, we extract the attention distribution from the most recently generated token to all preceding tokens via column-wise summation, i.e., $\mathbf{a}_{all}^{(t)} = \text{colsum}(\bar{\mathbf{A}}^{(t)}) \in \mathbb{R}^{P^{(t)}}$. Subsequently, we extract only the visual token attention scores from $\mathbf{a}_{all}^{(t)}$ using the image mask, denoted as $\mathbf{a}^{(t)} \in \mathbb{R}^{N_v}$, where N_v is the number of visual tokens. This visual attention vector effectively captures which visual tokens are most relevant to the current reasoning context.

As visual focus evolves across reasoning steps, we iteratively select the top-K attended visual tokens $\{\mathbf{v}_i^{(t)}\}_{i=1}^{K}$ as visual latents, which are integrated with hidden states $\mathbf{z}_t$. To prevent redundant re-selection and promote diverse exploration, we maintain a visited token set $\mathcal{V}_{\text{visited}}^{(t)}$, ensuring comprehensive visual coverage:

$$\mathcal{I}^{(t)} = \text{TopK}\left(\{a_i^{(t)} \mid i \notin \mathcal{V}_{\text{visited}}^{(t)}\}\right), \tag{5}$$

where $\mathcal{I}^{(t)}$ represents the indices of the K visual tokens with the highest attention scores at step t, and $\mathcal{V}_{\text{visited}}^{(t)}$ is updated after each selection. The original embeddings of the selected tokens $\mathbf{V}_{\mathcal{I}^{(t)}} = \{\mathbf{v}_i \mid i \in \mathcal{I}^{(t)}\}$ are further weighted by normalized attention scores:

$$\mathbf{B}^{(t)} = \text{Diag}(\text{Softmax}(\{a_i^{(t)} | i \in \mathcal{I}^{(t)}\}))\mathbf{V}_{\mathcal{I}^{(t)}}, \tag{6}$$

where $\text{Diag}(\cdot)$ constructs a diagonal matrix from a vector.

Spatially-Coherent Regularization. Although leveraging learned attention within Transformers provides explainable visual locations, it often suffers from limited spatial continuity due to the scattered nature of selected tokens and introduces noise associated with the attention sink phenomenon [86]. To mitigate these issues and enhance fine-grained perception [48] without external annotations, we introduce self-distillation supervision. This mechanism involves cropping the visual regions exhibiting spatial coherence, re-encoding them, and supervising the visual latents with these high-fidelity features.

Specifically, we first search for a $W \times W$ sub-grid patch that maximizes the density of attended visual tokens. Formally, we find the optimal top-left corner (r^*, c^*) within the valid grid bounds:

$$(r^*, c^*) = \underset{0 \le r, c \le G-W}{\operatorname{argmax}} \mathcal{N}(r, c), \tag{7}$$

where the density function $\mathcal{N}(r, c)$ counts the visited tokens falling within the window $\mathcal{N}(r, c) = \sum_{i \in \mathcal{I}^{(t)}} \mathbb{I}\big[r \le r_i < r + W, c \le c_i < c + W\big]$. Here, (r_i, c_i) denotes the row-column position of token i on the $G \times G$ grid, and $\mathbb{I}[\cdot]$ is the indicator function. This greedy selection ensures the cropped region captures a coherent visual context rather than scattered details. Then, the selected window is projected to pixel coordinates in the original image via mathematical transformation, cropped, and resized to the standard encoder input resolution using bilinear interpolation. We then re-encode this refined patch through the same visual encoder to obtain fine-grained representations $\{\mathbf{f}_i\}_{i=1}^{N_v'}$. A global pooling operation $\mathrm{Pool}(\cdot)$ is applied to yield a robust reference token $\mathbf{u}^{\text{ref}}$. Finally, we align the coarse-grained global token $\mathbf{b}^{(t)} = \mathrm{Pool}(\mathbf{B}^{(t)})$ with this high-fidelity reference using a lightweight token super-resolution module $\mathcal{F}_{\text{SR}} : \mathbb{R}^D \to \mathbb{R}^D$. We minimize the reconstruction error as follows:

$$\mathcal{L}_{\text{recon}}^{(t)} = \left\| \mathcal{F}_{\text{SR}} \left(\mathbf{b}^{(t)} \right) - \mathbf{u}^{\text{ref}} \right\|_2^2. \tag{8}$$

This supervisory signal enables the model to prioritize spatially coherent and semantically intact visual contexts during latent generation through self-distillation.

3.3 Routing Depth Scaling

While visual replay mechanisms significantly enhance fine-grained grounding by incorporating spatially-coherent visual latents, existing methods typically allocate uniform computational budgets across prefilling tokens and newly generated latents during contextual refinement. Intuitively, however, tokens and latents contribute heterogeneously to the final reasoning objective, rendering static budget allocation suboptimal. We hypothesize that a token's processing requirement is intrinsically linked to its relative *difficulty*, which can be quantified via optimization dynamics.

A Theoretical Perspective: Rank Dynamics as a Proxy for *Difficulty*. To formalize *difficulty*, we analyze the optimization dynamics of QKVO weight matrices, positing that their evolution reflects sample complexity. For simple samples, consistent gradient updates drive weights toward a compact, low-rank structure where principal singular values dominate and noise is suppressed. Conversely, difficult samples involve conflicting contexts that force representations to constantly adjust between competing patterns. This instability manifests as *resurrected* suppressed singular values or reshaped principal distributions, causing persistent fluctuation in the nuclear norm of matrix. We quantify these dynamics via the nuclear norm, defined as follows:

Definition 3.1 (Nuclear Norm of a Matrix)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be a weight matrix in transformer, the Singular Value Decomposition (SVD) of $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a diagonal matrix containing the non-negative singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$, with $r = \mathrm{rank}(\boldsymbol{A})$. The nuclear norm of $\boldsymbol{A}$, denoted as $\|\boldsymbol{A}\|_*$, is defined as the sum of its singular values:*

$$\|\boldsymbol{A}\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(\boldsymbol{A}) = \mathrm{Tr}\left(\sqrt{\boldsymbol{A}^\top \boldsymbol{A}}\right).$$

Based on above defined *difficulty* indicator, we reveal heterogeneous optimization complexities across tokens during latent training, termed the *fixed-depth optimization dilemma*. To address this without modifying the pretrained VLM architecture while effectively leveraging its inherent knowledge, we introduce a lightweight router that dynamically allocates additional reasoning steps exclusively to critical tokens, enabling adaptive reasoning depths.

Router Network. Let the input token sequence at the t-th reasoning step be denoted as $\mathbf{U}^{(t)} = [\mathbf{X} \| \mathbf{B}^{(1)} \| \mathbf{z}^{(1)} \| \cdots \| \mathbf{B}^{(t)} \| \mathbf{z}^{(t)}] \in \mathbb{R}^{P^{(t)} \times d}$, where $P^{(t)}$ represents the total sequence length and d is the hidden dimension. To facilitate the illustration of our depth scaling mechanism, we decompose the forward pass into layer-wise computations. Specifically, within the l-th transformer layer, we first compute the intermediate feature $\mathbf{H}^{(l,t)} = f(\mathbf{U}^{(t)}) \in \mathbb{R}^{P^{(t)} \times d}$, where $f(\cdot)$ denotes the transformation

of the l-th transformer layer in LLM. Subsequently, a lightweight router network computes a scalar importance score for each token based on its corresponding hidden representation:

$$\mathbf{s}^{(l,t)} = \mathcal{F}_{\text{router}}(\mathbf{H}^{(l,t)}) \in \mathbb{R}^{P^{(t)}}, \tag{9}$$

where the i-th element $s_i^{(l,t)}$ quantifies the importance of the i-th token in the l-th layer, and $\mathcal{F}_{\text{router}}^{(l)}(\cdot)$ denotes the independent router network in the l-th layer. We define $T_\alpha(\mathbf{s}^{(l,t)})$ as the index set of the top-α tokens with the highest scores, where α serves as a predefined hyper-parameter. In practice, α can be formulated as a layer-dependent function $\alpha(l)$ to enable adaptive resource allocation at different layers.

Depth Scaling Computation. Obtaining the router weight, we select the top-α tokens for depth scaling, i.e., repeat iteration in one transformer block. For a given depth scaling step $d \in \{1, \ldots, D\}$, the token-wise representation update rule within a transformer layer can be generally formulated as follows:

$$\mathbf{h}_i^{(d)} = \begin{cases} s_i^{(d)} \cdot f_{\in T_\alpha(\mathbf{s})}(\mathbf{h}_i^{(d-1)}, \mathbf{m}^{(d-1)}), & \text{if } i \in T_\alpha(\mathbf{s}^{(d)}), \\ \mathbf{h}_i^{(d-1)}, & \text{if } i \notin T_\alpha(\mathbf{s}^{(d)}). \end{cases} \tag{10}$$

For notational brevity, we omit the layer index l and reasoning step t, denoting refinement depth by d (e.g., $\mathbf{s}^{(l,t)} \to \mathbf{s}^{(d)}$). Let $\mathbf{h}_i^{(0)}$ denote the initial hidden state of token i, and $\mathbf{m}^{(d)}$ the attention mask at step d. A router network dynamically selects a critical token subset $T_\alpha(\mathbf{s}^{(d)})$ based on importance scores, acting as a binary gate. Consequently, the transformation $f(\cdot)$ is applied exclusively to tokens $i \in T_\alpha(\mathbf{s}^{(d)})$ for d refinement steps, while non-critical tokens retain their prior states to ensure efficiency. Finally, we aggregate the refined representations with step-aware positional encoding $\mathbf{e}^{(d)}$ to further form the final hidden state:

$$\mathbf{H}^{(D)} = \mathbf{H}^{(0)} + \sum_{d=1}^{D} \left(\mathbf{H}^{(d)} \odot \mathbf{e}^{(d)} \right). \tag{11}$$

This adaptive strategy allocates deeper reasoning pathways to critical tokens, enabling the model to capture complex visual contexts and facilitate more profound reasoning capabilities. We further provide a theoretical analysis demonstrating that, under mild assumptions, our design achieves a tighter generalization bound (see **Appendix**).

Theorem 3.1 (Generalization Analysis of Token-wise Depth Scaling)**.** *Let $\mathcal{T}_{std}$ denote the standard L-layer Transformer with total parameter count $W_{std} = \Theta(Ld^2)$. Let $\mathcal{T}_{topk}$ denote the token-wise depth scaling variant with Top-K selection, having total parameter count $W_{topk} = W_{std} + Md = W_{std}(1 + O(1/d))$, where $M = \lceil L/2 \rceil$ is the number of modified layers. Under Assumptions A.1–A.6, with probability at least $1 - \delta$, the population risk satisfies:*

$$\mathcal{L}_{\mathcal{D}}(f_{topk}) \leq \mathcal{L}_{\mathcal{D}}(f_{std}) - \frac{\delta_0 KL}{4n} \cdot \eta_{eff} + \tilde{\mathcal{O}}\left(\sqrt{\frac{dL \log(nL/\delta)}{N}} \right) - \Omega\left(\frac{K^2 L}{n^2} \sqrt{\frac{W_{std}}{N}} \right), \tag{12}$$

where $\delta_0 > 0$ is the minimal improvement from single-layer iteration, $K = \rho n$ is the selection number, η_{eff} is the effective selection rate, d is the hidden dimension, and N is the number of training samples.

3.4 Training Procedure

Curriculum Latent Training. To reduce annotation overhead and avoid human priors limiting learning, we do not directly supervise intermediate latent representations. Instead, we introduce a curriculum that links implicit latents with explicit reasoning chains. We first train the model with standard Chain-of-Thought (CoT) supervision, generating all reasoning steps explicitly to build baseline reasoning ability. Then, over training, we gradually replace one explicit step at a time with an informative ⟨latent⟩ token. This curriculum enables each latent to progressively ground relevant contextual cues, internalizing explicit reasoning into compact latent representations.

Training Objective. The overall training objective combines the standard language modeling loss with the self-distillation loss in the SCF-VR module,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \frac{1}{T_r} \sum_{t=1}^{T_r} \mathcal{L}_{\text{recon}}^{(t)}, \quad \mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T_r+T_a} \log p_\theta \left(a_t | \mathbf{U}^{(t)}, a_{<t} \right), \tag{13}$$

Table 1: Comparison of various multimodal reasoning baselines across three benchmarks. We selected three datasets featuring detailed reasoning chains for training and evaluate on their test split, respectively. Three metrics are reported: Accuracy (%), Average number of Autoregressive Steps (# AR steps), and Average Generation Time (AVG. Time). † denotes the reimplementation for the methods with the same configuration with ours. No-CoT notes that directly predicts answers without generating intermediate steps.

		M^{3}CoT			ScienceQA			GQA		
Method	**Model**	Acc.(%) ↑	# AR Steps ↓	Avg. Time ↓	Acc.(%) ↑	# AR Steps ↓	Avg. Time ↓	Acc.(%) ↑	# AR Steps ↓	Avg. Time ↓
No-CoT	Qwen2-VL 7B	45.4	–	–	64.4	–	–	–	–	–
Multimodal CoT [108]		42.5	106.3	3.10	58.3	83.9	2.44	–	–	–
CCoT [49]		44.1	177.2	5.31	63.8	164.0	5.23	51.2	76.4	7.21
ICoT [15]		46.0	96.5	2.86	65.4	77.4	2.28	–	–	–
SCAFFOLD [30]		44.9	170.8	5.14	62.5	162.3	4.91	48.7	72.8	6.72
Chain-of-Focus [107]		64.3	185.7	2.63	91.2	162.3	2.09	61.8	128.6	3.01
IVT-LR†		69.8	10.0	0.67	92.8	11.0	0.81	65.8	10.1	0.68
Ours		**73.0**	7.0	0.86	**95.9**	7.2	1.02	**67.4**	9.2	0.82
No-CoT	Chameleon 7B	28.4	–	–	48.5	–	–	–	–	–
Multimodal CoT [108]		30.6	110.5	3.62	50.7	98.7	3.33	–	–	–
CCoT [49]		31.4	168.4	5.35	51.3	174.2	5.39	33.1	150.6	5.31
ICoT [15]		32.3	110.9	5.43	53.4	92.4	4.62	–	–	–
SCAFFOLD [30]		31.1	194.3	6.12	47.5	160.6	6.03	32.8	156.0	4.17
Chain-of-Focus [107]		36.5	739.4	3.09	61.2	717.1	2.56	34.6	360.4	2.98
IVT-LR†		40.8	10.0	1.13	63.2	11.0	1.56	38.1	10.1	0.98
Ours		**43.4**	7.0	1.24	**65.7**	7.2	1.37	**39.4**	9.2	1.21

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss over the language modeling task, where λ is a hyper-parameter balancing the primary task and the auxiliary reconstruction objective. During inference, only the LLM component is active, and the visual play module is disabled, ensuring no additional computational overhead at test time.

Table 2: Performance comparison of our method against baselines across four paradigms. Benchmarks are categorized into three domains: **Visual Perception**, **Compositional Reasoning**, and **Mathematical Reasoning**. Among latent reasoning approaches, the best results are highlighted in **bold**, and the second-best are underlined. Our method is built upon the Qwen2.5-VL-7B backbone.

Model	**Data Size**	**Vision-centric Reasoning**				**Compositional**		**Mathematics Reasoning**		
		MMVP	SeedBench-2-Plus	HallusionBench	HRBench	BLINK	MMStar	MathVista	MathVision	MM-Math
Zero-Shot VLMs										
GPT-4o [51]	–	68.7	72.0	–	–	68.0	64.7	63.8	30.4	31.8
Qwen2.5-VL-7B [2]	–	65.7	65.3	56.6	68.3	53.6	59.7	68.2	25.6	37.5
LLaVA-OneVision [33]	*9M*	74.0	61.2	51.1	63.0	49.3	59.1	58.6	-	-
InternVL3.5-8B [77]	*70K*	57.7	69.8	56.2	59.4	54.8	53.3	71.6	28.3	-
Explicit CoT Reasoning										
Multimodal CoT [108]	–	68.1	54.1	63.6	-	-	57.9	56.4	21.8	35.6
CCoT [49]	–	69.0	69.0	64.9	-	-	58.7	57.8	22.5	36.3
ICoT [15]	–	69.3	70.3	65.5	-	-	60.4	58.9	23.3	37.0
Tool-use & RL Enhanced Reasoning										
PAPO [82]	*39K*	68.7	54.1	57.5	68.1	52.7	45.8	67.5	-	-
Vision-R1 [27]	*200K*	72.7	69.0	63.8	75.1	52.7	62.7	52.4	-	40.2
VL-Rethinker [71]	*39K*	72.7	70.3	71.1	63.5	55.6	63.2	72.8	29.3	-
DeepEyes [112]	*47K*	70.0	69.1	62.6	69.1	51.1	58.7	70.1	26.6	-
Latent Reasoning										
LVR [32]	*470K*	64.0	47.4	65.2	53.6	53.6	57.9	-	-	-
Monet [75]	*125K*	68.0	65.9	56.4	68.0	50.7	60.3	-	-	-
DMLR [40]	–	70.1	-	65.8	-	56.9	60.3	59.1	24.4	38.8
Laser [80]	*267K*	72.0	**70.1**	67.7	72.5	-	60.1	-	-	-
Ours	*30K*	**76.7**	66.9	**68.6**	**74.2**	**58.0**	**60.8**	**69.8**	**25.9**	**39.8**

4 Experiments

4.1 Experimental Setup

Evaluation Benchmark. We evaluate our method on three tasks across twelve benchmarks: (1) Mathematics Reasoning (MathVista [44], MathVision [73], MM-Math [61]); (2) Vision-centric Reasoning (Hallusion-Bench [16], MMVP [69], Seed-Bench-2-Plus [34], HR-Bench [78], GQA [28]); (3) Multimodal Composition Reasoning (MMStar [5], BLINK [13], ScienceQA [45], M^{3}CoT [6]).

Implementation Details. All models use eager attention for explicit access to internal attention maps. We set latent reasoning tokens to $T_r = 4$, inject $\alpha = 32$ salient regions per refinement, and

Table 3: We perform comprehensive ablation studies to analyze the contribution of each design choice. Experiments are conducted on three CoT benchmarks with multiple backbone models. To save the training overhead, all models are trained for 16 epochs with a balanced sampled subset.

Method	Model	M^{3}CoT ↑	ScienceQA ↑	GQA ↑
Base	Qwen2-VL 2B	44.2	62.4	38.7
+RDS		47.6 (+3.4)	64.9 (+2.5)	40.4 (+1.7)
+RDS & SCF-VR		51.2 (+7.0)	66.3 (+3.9)	40.9 (+2.2)
Base	Qwen2-VL 7B	63.7	78.2	52.3
+RDS		65.2 (+2.5)	80.5 (+2.3)	53.7 (+1.4)
+RDS & SCF-VR		66.4 (+3.6)	81.3 (+3.1)	54.1 (+1.8)
Base	Qwen2.5-VL 3B	62.9	76.5	51.9
+RDS		64.5 (+1.6)	77.8 (+1.3)	52.9 (+1.0)
+RDS & SCF-VR		65.3 (+2.4)	78.5 (+2.0)	53.6 (+1.7)
Base	Qwen2.5-VL 7B	71.1	85.7	56.3
+RDS		72.9 (+1.8)	87.2 (+1.5)	57.6 (+1.3)
+RDS & SCF-VR		73.9 (+2.8)	88.2 (+2.5)	58.3 (+2.0)

Table 4: **Ablation Experiments.** We provide ablation analysis of key parameters and experimental settings on three benchmarks. All the variants adopt Qwen2.5-VL-3B as the base model. **Base** refers the model that adopts RDS and SCF-VR.

Dataset	M^{3}CoT	ScienceQA	GQA
(a) Curriculum Training			
Base	62.0	75.4	51.3
Base + Curr.	62.9 (+0.9)	76.5 (+1.1)	51.9 (+0.6)
(b) Token Selection in Router $T_\alpha(\mathbf{s})$			
$\alpha = 32$	62.9	76.5	51.9
$\alpha = 16$	62.1 (-0.8)	75.3 (-1.2)	51.2 (-0.7)
$\alpha = 64$	62.8 (-0.1)	76.2 (-0.3)	51.9 (-0.0)
Cosine-annealed Retention	62.6 (-0.3)	76.5 (-0.0)	51.7 (-0.2)
(c) Selection Strategy			
32-Patch	62.9	76.5	51.9
16-Patch	61.7 (-1.2)	75.5 (-1.0)	51.3 (-0.6)
Soft-Mix	62.2 (-0.7)	75.8 (-0.7)	51.9 (-0.0)
(d) Position Encoding			
Keep Old Position	62.9	76.5	51.9
Rearrange Position	62.2 (-0.7)	76.5 (-0.0)	51.3 (-0.6)

cap refinement depth at $D = 1$ for efficiency. The region-injection count follows cosine annealing from $\alpha_s = 64$ to $\alpha_e = 16$, and training runs for 16 epochs by default. We optimize with DeepSpeed ZeRO-2 (no CPU offloading), using batch size 8 per GPU and Adam with learning rate 4×10^{-5} and $\beta_1 = 0.9$. For self-distillation, we set $\lambda = 1.0$ for 2B/3B models and $\lambda = 0.2$ for 7B models. Images are resized according to the training dataset. Experiments are conducted on 16 NVIDIA H20 GPUs (96GB each). See the **Appendix** for more details.

4.2 Overall Quantitative Results

Reasoning Accuracy. As summarized in Table 1, trained with identical protocols, our method achieves the best results on three fine-grained CoT datasets across all backbones. Compared with explicit CoT baselines, it brings an average accuracy gain of nearly 30% on Qwen-2-VL. It also improves over the strongest latent baseline (IVT-LR) by an average of +2.6% across all benchmarks. As shown in Table 2, we further evaluate on mainstream diagnostic benchmarks. We observe consistent improvements on most tasks, with the largest gain on MMVP (+4.7%), which requires fine-grained visual discrimination. These improvements are attributed to our SCF-VR, which enforces spatially coherent visual latents to reduce hallucination while preserving visual details. In contrast, gains are more modest on text-intensive benchmarks (e.g., SeedBench-2-Plus), likely due to limited domain-specific data. Notably, despite operating purely in the latent space without external retrieval or RL-based optimization, our approach achieves competitive performance against explicit CoT and tool-augmented methods through token-wise depth scaling.

Reasoning Efficiency. Beyond accuracy, our method significantly improves inference efficiency, measured by (1) **autoregressive steps** and (2) **wall-clock latency**. **(1) Autoregressive Steps.** Across all backbones, we reduce autoregressive generation steps by at least $10\times$ versus traditional explicit CoT baselines. This stems from performing compact multi-step reasoning in the latent space, avoiding verbose textual rationales required by explicit CoT. **(2) Inference Latency.** On the Qwen backbone, our method achieves $\sim$0.9s average latency, comparable to IVT-IR, while generating rationales 3–6$\times$ faster than explicit CoT competitors. Similar trends hold for Chameleon. Although the No-CoT baseline is fastest (0.35s), it lacks strong multi-step reasoning. Overall, our approach reaches near this efficiency frontier, delivering state-of-the-art accuracy with only marginal latency overhead, enabling fast yet nuanced multimodal reasoning.

4.3 Ablation Analysis

Effect of Designed SCF-VR and RDS. We introduce several model variants to validate the effectiveness of our proposed modules across four representative MLLM backbones. As summarized in Table 3, consistent performance improvements are observed across all architectures. Notably, when instantiated on Qwen2-VL-2B, our method achieves a substantial improvement of +7.0% on M^{3}CoT. Furthermore, we observe that performance gains are less pronounced on the GQA benchmark. We hypothesize that this trend arises from the relatively straightforward nature of GQA, where both visual

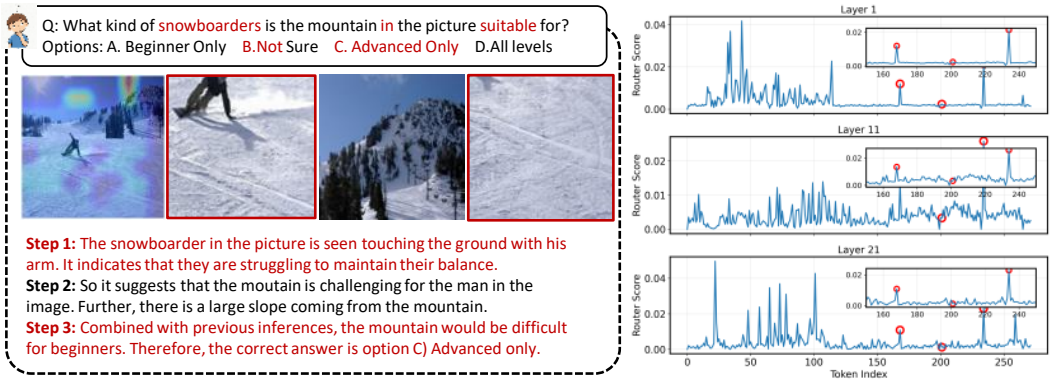
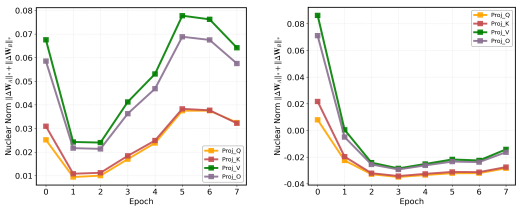


Figure 3: Visualization of learned router scores and selected top-K tokens for depth scaling, revealing which textual tokens, CoT latent tokens, and visual regions exhibit greater optimization difficulty.

inputs and queries demand less complex reasoning. Consequently, baseline models can adequately address the task requirements, rendering the contributions of our designed modules less impactful in such scenarios.

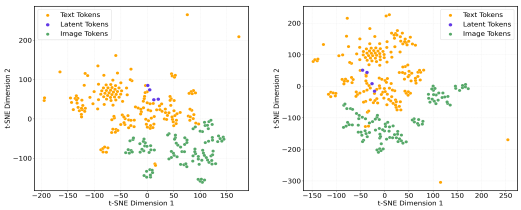
Routing Behavior. As illustrated in Fig. 3, the left panel highlights tokens selected by the learnable router (marked in red color). In our latent space reasoning framework, each reasoning step is compressed into a single latent token. The right panel displays the router scores for these latent tokens (circled in red), where the first and third latent token exhibit significantly higher scores, identifying them as *difficult* tokens requiring additional depth scaling. Specifically, the first reasoning step analyzes critical visual details essential for the answer, such as the skier's arms touching the ground for balance, which indicates the difficulty level of the ski slope. Furthermore, the visualized attention scores in image reveal that the router prioritizes regions corresponding to snow, body limbs, and trees. This demonstrates that the model achieves deeper compositional reasoning by iteratively refining these semantic hubs.

5 In-depth Analysis



(a) Without RDS (b) With RDS

Figure 4: (a) and (b) illustrate the Nuclear norm variance without RDS and with RDS.



(a) Vanilla (b) Ours

Figure 5: (a) and (b) illustrate the t-SNE distribution of the learned latents, visual tokens, and text tokens.

Visualization of Crop Region. As illustrated in Fig. 6, we visualize the attention-guided crops across successive reasoning steps. Taking the right-upper case as a representative example, which queries the purpose of the train cart, the model initially localizes the region corresponding to the description "shape like a house" during the first reasoning step. In subsequent steps, the attention focus progressively shifts toward the "entrance door". Remarkably, without any explicit fine-grained supervision, our curriculum training paradigm enables latent tokens to progressively capture logical dependencies and establish robust visual-semantic alignments. This empirically demonstrates the model's capacity to jointly perform multi-step visual grounding and generate coherent reasoning chains.

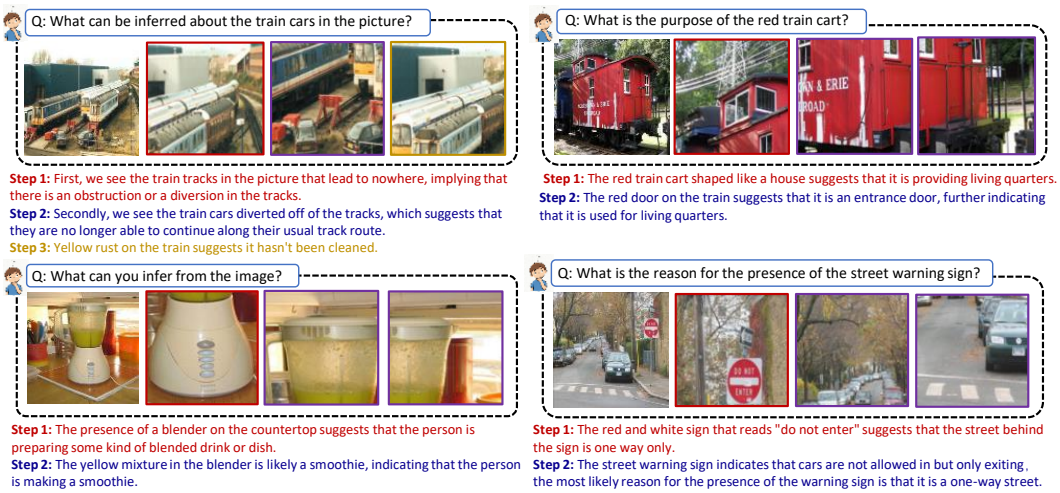


Figure 6: Visualization of cropped region in each latent reasoning step.

Nuclear Norm Behavior with RDS. As illustrated in Fig.4a and Fig.4b, we conducted training experiments on a subset of hard samples. The results demonstrate that after incorporating the RDS module, the cumulative nuclear norm of the QKVO projection weights achieves stable convergence by the 3rd epoch. This observation suggests that the depth scaling strategy effectively mitigates gradient volatility and uncertainty induced by hard samples, thereby substantially enhancing the model's capacity to learn and fit challenging instances and critical difficult tokens.

Latent Behavior Analysis. As depicted in Fig. 5a and Fig. 5b, we visualize the embedding space distribution of multimodal features for both the baseline and our method. The latent tokens learned by our approach form several distinct clusters, situated centrally within the text embedding manifold. Notably, compared to the baseline, our latents exhibit closer proximity to visual embeddings. This observation suggests that our latent tokens encapsulate richer reasoning semantics while facilitating deeper integration of visual information.

6 Conclusions and Limitations

In this paper, we empirically reveal two critical observations: the vision-text optimization disparity and the fixed-depth optimization dilemma. In light of these findings, we propose a visual replay module and routing depth scaling to collaboratively enhance visual perception and exploit critical tokens for deeper contextual reasoning. Instead of relying on extensive latent-supervised annotations or predefined visual priors, we adopt curriculum training to progressively capture rich contextual information, yielding highly informative latent representations. Extensive experiments demonstrate that: (1) our approach not only reduces decoding latency but also achieves superior performance across diverse benchmarks; (2) the proposed components exhibit strong generalization across various backbones and tasks; and (3) visualization analysis reveals both quantitative and qualitative gains afforded by enhanced visual grounding and adaptive token depth scaling. In future work, we plan to extend this framework to more complex reasoning scenarios, such as long-term video understanding, and scale it to larger model architectures.

While our approach demonstrates substantial empirical gains, the adaptive routing mechanism introduces mild variability in inference latency, which may necessitate careful scheduling for real-time applications with strict timing constraints. Moreover, we adopt a data-driven adaptive strategy to learn the router for token difficulty assessment, eliminating the need for supervisory feedback from external expert models. Our theoretical analysis suggests that routing accuracy plays an important role in the generalization performance of latent reasoning.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (62302045), the Fundamental Research Funds for the Central Universities, and in part by the Opening Project of the State Key Laboratory of General Artificial Intelligence, BIGAI/Peking University, Beijing, China Project No. SKLAGI20250P05.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. 2025. URL https://arxiv.org/abs/2511.21631. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8, 31, 32

[3] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *CoRR*, abs/2505.14231, 2025. 4

[4] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930, 2024. 3

[5] 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. 8, 31, 34

[6] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proc. of ACL*, 2024. 8, 31

[7] Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024. 4

[8] Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025. 3

[9] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. 1

[10] Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9199–9209. Computer Vision Foundation / IEEE, 2025. 3

[11] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948. 1, 3

[12] Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. Onethinker: All-in-one reasoning model for image and video. *CoRR*, abs/2512.03043, 2025. 31

[13] 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*, pages 148–166. Springer, 2024. 8, 31, 34

[14] Chang Gao, Chujie Zheng, Xiong hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *ArXiv*, abs/2511.20347, 2025. URL https://api.semanticscholar.org/CorpusID:283251588. 1

[15] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025. doi: 10.1109/CVPR52734.2025.01818. 3, 8, 32

[16] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 8, 31, 34

[17] Yudong Han, Jianhua Yin, Jianlong Wu, Yinwei Wei, and Liqiang Nie. Semantic-aware modular capsule routing for visual question answering. *IEEE Trans. Image Process.*, 32: 5537–5549, 2023. 3

[18] Yudong Han, Yupeng Hu, Xuemeng Song, Haoyu Tang, Mingzhu Xu, and Liqiang Nie. Exploiting the social-like prior in transformer for visual reasoning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 2058–2066. AAAI Press, 2024.

[19] Yudong Han, Qingpei Guo, Liyuan Pan, Liu Liu, Yu Guan, and Ming Yang. Dynfocus: Dynamic cooperative network empowers llms with video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8512–8522. Computer Vision Foundation / IEEE, 2025. 1, 3

[20] Yudong Han, Yong Wang, Zaiquan Yang, Liang Lin, Chongyang Tao, Xiangxiang Chu, and Liyuan Pan. Beyond entropy: Self-diagnostic multi-role token optimization for video reasoning. 2026. URL https://arxiv.org/abs/2610.03400. 3, 4

[21] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, et al. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024. 2, 3

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9. 2

[23] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022. 3

[24] 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. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 139348–139379, 2024. 3

[25] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, August 2024. 3

[26] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749, 2025. 3

[27] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 8, 32, 33

[28] 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*, pages 6700–6709, 2019. 8, 31

[29] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. 1

[30] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2886–2903, 2025. 1, 8, 32

[31] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025. 4, 33

[32] Bangzheng Li et al. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025. 8, 32

[33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2024. 8, 32

[34] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench-2-Plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 8, 31, 34

[35] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 3

[36] Xiangqi Li, Libo Huang, Zhulin An, Weilun Feng, Chuanguang Yang, Boyu Diao, Fei Wang, and Yongjun Xu. Geometric feature embedding for effective 3d few-shot class incremental learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 34674–34687. PMLR, 2025. 2

[37] Xiangqi Li, Libo Huang, Jiarui Zhao, Weilun Feng, Chuanguang Yang, Zhulin An, and Yongjun Xu. Point-uq: An uncertainty-quantification paradigm for point cloud few-shot class incremental learning. In *International Conference on Learning Representations*, volume 2026, pages 17529–17551, 2026. 2

[38] Xiangqi Li, Libo Huang, Jiarui Zhao, Weilun Feng, Chuanguang Yang, Zhulin An, and Yongjun Xu. Scenescaffold: Active scene-state construction for unified 3d scene understanding. In *Advances in Neural Information Processing Systems*, 2026. 3

[39] Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 48–58, October 2025. 4

[40] Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space, 2025. URL https://arxiv.org/abs/2512.12623. 4, 8, 32, 33

[41] Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. Visual abstract thinking empowers multimodal reasoning. *arXiv preprint arXiv:2505.20164*, 2025. 3

[42] Ying Liu, Yudong Han, Kean Shi, and Liyuan Pan. Mema: Memory-augmented adapter for enhanced vision-language understanding. *arXiv preprint arXiv:2603.00655*, 2026. 3

[43] Ying Liu, Yudong Han, Kean Shi, and Liyuan Pan. Mema: Memory-augmented adapter for enhanced vision-language understanding. 2026. URL https://arxiv.org/abs/2603.00655. 4

[44] 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 *The Twelfth International Conference on Learning Representations*. 8, 31, 34

[45] Pan Lu, Swaroop Mishra, Tony 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. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 2507–2521, 2022. 8, 31

[46] Haipeng Luo, Huawen Feng, Qingfeng Sun, Can Xu, Kai Zheng, Yufei Wang, Tao Yang, Winston Hu, and Yansong Tang. Agentmath: Empowering mathematical reasoning for large language models via tool-augmented agent. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. 1

[47] Weijian Ma, Shizhao Sun, Tianyu Yu, Ruiyu Wang, Tat-Seng Chua, and Jiang Bian. Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation. *CoRR*, abs/2601.01984, 2026. 3

[48] Ziyu Ma, Shidong Yang, Yuxiang Ji, Yiming Hu, Tongwen Huang, Yong Wang, Jianfei Cai, and Xiangxiang Chu. Towards high-resolution visual perception via hierarchical entity exploration. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari, editors, *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part XXIV*, volume 17024 of *Lecture Notes in Computer Science*, pages 533–550. Springer, 2026. 5

[49] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. doi: 10.1109/CVPR52733.2024.01367. 1, 3, 8, 32, 33

[50] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806, 2024. doi: 10.1609/aaai.v38i17.29844. 3

[51] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8, 32

[52] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8228–8237. IEEE, 2022. 4

[53] Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025. 4

[54] Zhen Qu, Xian Tao, Xiaoyi Bao, Dingrong Wang, Shichen Qu, Zhengtao Zhang, and Xingang Wang. AG-VAS: anchor-guided zero-shot visual anomaly segmentation with large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pages 14126–14136. Computer Vision Foundation, 2026. 1

[55] Matthew Renze and Erhan Guven. Self-reflection in LLM agents: Effects on problem-solving performance. *CoRR*, abs/2405.06682, 2024. 3

[56] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 8612–8642, 2024. 3

[57] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024. 1

[58] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: compressing chain-of-thought into continuous space via self-distillation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 677–693. Association for Computational Linguistics, 2025. 2

[59] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025. 4

[60] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *ArXiv*, abs/2504.07951, 2025. 1

[61] Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juan-Zi Li. MM-MATH: advancing multimodal math evaluation with process evaluation and fine-grained classification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 1358–1375. Association for Computational Linguistics, 2024. 8, 31, 34

[62] Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Ruihua Song. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. *CoRR*, abs/2505.16552, 2025. 2

[63] Xiaole Tang, Xiaoyi He, Jiayi Xu, Xiang Gu, and Jian Sun. Learning continuous wasserstein barycenter space for generalized all-in-one image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[64] Xiaole Tang, Jiayi Xu, Xiang Gu, Yan Yang, and Jian Sun. Binding multiple modalities via multimodal wasserstein barycenter. *arXiv preprint arXiv:2609.33800*, 2026. 2

[65] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 31

[66] Kwai Keye Team, Bin Wen, Changyi Liu, Chengru Song, Chongling Rao, Guowang Zhang, Han Li, Haonan Fan, Hengrui Ju, Jiankang Chen, Jiapeng Chen, Jiawei Yuan, Kaixuan Yang, Kaiyu Jiang, Kun Gai, Lingzhi Zhou, Na Nie, Sen Na, Tianke Zhang, Tingting Gao, Xuanyu Zheng, Yulong Chen, Fan Yang, Haixuan Gao, Lele Yang, Mingqiao Liu, Muxi Diao, Qi Zhang, Qile Su, Wei Chen, Wentao Hong, Xingyu Lu, Yancheng Long, Yankai Yang, Yingxin Li, Yiyang Fan, Yu Xia, Yuzhe Chen, Ziliang Lai, Chuan Yi, Haonan Jia, Tianming Liang, Weixin Xu, Xiaoxiao Ma, Yang Tian, Yufei Han, Feng Han, Hang Li, Jing Wang, Jinghui Jia, Junmin Chen, Junyu Shi, and Ruilin Zhang. Kwai keye-vl-2.0 technical report. 2026. URL https://arxiv.org/abs/2606.10651. 1

[67] Yang Tian, Fan Liu, Jingyuan Zhang, Wei Bi, Yupeng Hu, and Liqiang Nie. Open multimodal retrieval-augmented factual image generation. *arXiv preprint arXiv:2510.22521*, 2025. 3

[68] Yang Tian, Fan Liu, Jingyuan Zhang, Yupeng Hu, Liqiang Nie, et al. Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32967–32982, 2025. 3

[69] 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*, pages 9568–9578, 2024. 8, 31, 34

[70] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *CoRR*, abs/2504.08837, 2025. 3

[71] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025. 8, 32, 33

[72] Junjian Wang, Xin Zhou, Qiran Xu, and Kun Zhan. M2A: synergizing mathematical and agentic reasoning in large language models. *CoRR*, abs/2605.09879, 2026. 1

[73] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 8, 31, 34

[74] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 31

[75] Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*, 2025. 8, 32, 33

[76] Weiyun Wang, Zhangwei Gao, Zhe Chen, et al. Internvl 3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 32

[77] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 8

[78] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025. 8, 31, 34

[79] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 50208–50232. PMLR / OpenReview.net, 2024. 1

[80] Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. Forest before trees: Latent superposition for efficient visual reasoning. *arXiv preprint arXiv:2601.06803*, 2026. 4, 8, 32, 33

[81] Zengbin Wang, Feng Xiong, Liang Lin, Xuecai Hu, Yong Wang, Yanlin Wang, Man Zhang, and Xiangxiang Chu. Visually-guided policy optimization for multimodal reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 6646–6664. Association for Computational Linguistics, 2026. 2

[82] 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. 8, 32, 33

[83] Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 2250–2259. IEEE, 2025. 2

[84] Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. *CoRR*, abs/2507.10203, 2025. 4

[85] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. *CoRR*, abs/2509.20317, 2025. 4

[86] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 5

[87] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025. 3

[88] Yibo Yan, Shen Wang, Jiahao Huo, Philip S. Yu, Xuming Hu, and Qingsong Wen. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. In Georg Rehm and Yunyao Li, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 69–82. Association for Computational Linguistics, 2025. 1

[89] Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24902–24912, June 2024. doi: 10.1109/CVPR52733.2024.02352. 1

[90] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1

[91] Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in MLLM reasoning. *CoRR*, abs/2507.03019, 2025. 2

[92] Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in mllm reasoning. 2025. 1, 3

[93] Zaiquan Yang, Yuhao Liu, Jiaying Lin, Gerhard Hancke, and Rynson W Lau. Boosting weakly supervised referring image segmentation via progressive comprehension. *Advances in Neural Information Processing Systems*, 37:93213–93239, 2024. 3

[94] Zaiquan Yang, Yuhao Liu, Gerhard Hancke, and Rynson WH Lau. Unleashing the potential of multimodal llms for zero-shot spatio-temporal video grounding. *arXiv preprint arXiv:2509.15178*, 2025. 3

[95] Zaiquan Yang, Fei Wei, Yong Wang, Yudong Han, Yiyu Li, Zhuofan Zong, Gerhard Petrus Hancke, Xiangxiang Chu, and Rynson WH Lau. Towards unbiased on-policy distillation for block diffusion language models. 2026. URL https://arxiv.org/abs/2610.05373. 3

[96] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025. 4

[97] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Honglin Yu, Weinan Dai, Yuxuan Song, Xiang Wei, Haodong Zhou,

Jingjing Liu, Wei Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yong-Xu Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025. URL https://api.semanticscholar.org/CorpusID:277104124. 1

[98] Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. *CoRR*, abs/2502.17425, 2025. 1

[99] Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. *ArXiv*, abs/2511.11007, 2025. URL https://api.semanticscholar.org/CorpusID:283055301. 3, 4

[100] Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, Guibin Zhang, Jiale Tao, Jiayi Zhang, Siyuan Ma, Kaituo Feng, Haojie Huang, Youxing Li, Rong Chen, Huacan Wang, Chenglin Wu, Z.Y. Su, Xiaogang Xu, Kelu Yao, Kun Wang, Chen Gao, Yue Liao, Ruqi Huang, Tao Jin, Cheng Tan, Jiang-She Zhang, Wenqi Ren, Yanwei Fu, Yong Liu, Yu Wang, Xiangyu Yue, Yu-Gang Jiang, and Shuicheng Yan. The latent space: Foundation, evolution, mechanism, ability, and outlook. abs/2604.02029, 2026. URL https://api.semanticscholar.org/CorpusID:287074162. 3

[101] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2

[102] Rongchao Zhang, Yu Huang, Yongzhi Cao, and Hanpin Wang. Molebridge: Synthetic space projecting with discrete markov bridges. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. 1

[103] Rongchao Zhang, Weiping Ding, Hongbin Han, Yongzhi Cao, Hanpin Wang, and Yu Huang. Strfilter: Multi-modal medical image fusion via structure-oriented adaptive filtering. *Inf. Fusion*, 128:103888, 2026. 2

[104] Rongchao Zhang, Chengxin Li, Yiwei Lou, Yuling Shi, Hanpin Wang, and Yu Huang. Steering where to diffuse: Generative modeling of phenotypic response simulation with steered diffusion bridge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pages 27367–27377. Computer Vision Foundation, 2026. 1

[105] Rongchao Zhang, Yiwei Lou, Yu Huang, Yi Xin, Yongzhi Cao, and Hanpin Wang. Beyond conservation: Flexible molecular assembly with unbalanced diffusion bridge. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 28364–28372. AAAI Press, 2026. 1

[106] Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information flow in lvlms across reasoning tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 2289–2299. Association for Computational Linguistics, 2025. 4

[107] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025. 3, 8, 32, 33

[108] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024, 2024. 1, 2, 3, 8, 32, 33

[109] Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *CoRR*, abs/2601.18734, 2026. 3

[110] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 5168–5191, 2023. 3

[111] Tong Zheng, Xidong Wu, Zheng Zhang, Zhankui He, Chaoyi Zhang, Benjamin Coleman, Ruoqiao Wei, Di Bai, Haolin Liu, Rui Liu, Xue Wang, Yue Zhuan, Wang-Cheng Kang, Renkai Xiang, Heng Huang, Xinwu Cheng, and Yunsong Guo. Dream-rsi: Recursive self-improvement through evolving worlds. 2026. URL https://arxiv.org/abs/2609.14858. 1

[112] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 8, 32, 33

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In Sec.1, we make the claims to reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Sec. 6, we discuss the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in Sec. B and the Appendix for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will consider releasing the data and code once the paper is accepted.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide details about the experimental setting (*e.g.*, hyper-parameters) in Sec. B and the Appendix for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars due to limited computing resources.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information on the computer resources in Sec. B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We believe that its impact depends on how it is used and promoted.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the existing assets.

    Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Some LLMs are acted as our baselines in our experiments. We clarify their settings and usage following open source benchmark.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Abstract of Appendix

This appendix provides the **Proof of Theorem** in Eq. 12 (Appendix A), **More Experimental Details** (Appendix B), **Baseline Details** (Appendix C), **Benchmark Details** (Appendix D), and **Parametric Analysis** (Appendix E).

A Proof of Theorem 3.1

A.1 Problem Setup and Notation: What Are We Analyzing?

A.1.1 The Two Architectures

Standard Transformer ($\mathcal{T}_{\text{std}}$). Every token goes through exactly L layers. Each layer applies self-attention over *all* n tokens and FFN transformations.

Token-wise Depth Scaling ($\mathcal{T}_{\text{topk}}$).

- Odd layers $(1, 3, 5, \ldots)$: "Modified layers" where we can apply *double computation* to some tokens
- Even layers $(2, 4, 6, \ldots)$: Standard layers, same as $\mathcal{T}_{\text{std}}$
- In each modified layer: route network selects $K = \rho n$ "hard" tokens for a second pass through the *same* layer parameters. Notbly, we don't add new parameters for the second pass—we reuse the same weights. Only the routing network adds Md extra parameters where $M = L/2$, which is negligible ($< 0.06\%$ for $d \geq 768$).

A.2 Why This Might Help: Two Intuitions

Adaptive Compute Allocation. Not all tokens are equally difficult. A token like "the" in "the cat sat" is easy; a token requiring multi-hop reasoning is hard. By applying extra computation only to hard tokens, we use our parameter budget more efficiently.

Sparse Attention Benefit. When we recompute only K tokens, their attention is $K \times K$ instead of $n \times n$. This reduces the "effective Lipschitz constant" of the transformation, improving generalization.

A.3 Assumptions: What Do We Need?

Assumption A.1 (Weight Boundedness)**.** *All weight matrices have bounded operator norm:* $\|W_Q^{(l)}\|_{op}, \|W_K^{(l)}\|_{op}, \|W_V^{(l)}\|_{op}, \|W_O^{(l)}\|_{op} \leq B$, *and* $\|W_1^{(l)}\|_{op} \leq B_1, \|W_2^{(l)}\|_{op} \leq B_2$ *with* $B_1 B_2 \leq B^2$, *plus* $\|w_g^{(l)}\|_2 \leq B_g$ *for routing, where* $\|W\|_{op} = \sigma_{max}(W)$.

Assumption A.2 (Bounded Representations)**.** $\|X^{(1)}\|_F \leq R$, *and all intermediate representations satisfy* $\|\tilde{X}^{(l)}\|_F \leq R_{\text{ln}}, \|\hat{X}^{(l)}_{S^{(l)}}\|_F \leq R_{\text{ln}}$. *LayerNorm keeps things bounded automatically.*

Assumption A.3 (Iterative Improvement Exists)**.** *There exists a set* $\mathcal{G}$ *of "improvable"* $(\tilde{x}, y)$ *pairs where second-pass computation helps:*

$$\mathcal{G} := \{(\tilde{x}, y) : \mathbb{E}[\ell(\hat{x}, y) - \ell(\tilde{x}, y) \mid \tilde{x}, y] \leq -\delta_0\}, \tag{14}$$

with $\gamma_{\text{eff}} := \mathbb{P}[(\tilde{x}, y) \in \mathcal{G}] > 0$ *and stability under small perturbations. This is our* weakest *and most crucial assumption. We only assume: (1) some tokens can benefit from extra compute, and (2) this property is stable. We do NOT assume convexity or any specific loss shape.*

Assumption A.4 (Oracle Routing Exists)**.** *There exists an ideal (possibly unknown) linear classifier* (w_g^*, τ^*) *that can identify improvable tokens with accuracy* $\geq 1 - \epsilon_{\text{oracle}}$.

Assumption A.5 (Learned Routing Approximates Oracle)**.** *The trained routing satisfies* $\mathbb{E}[|S^{(l)} \triangle S^*|/K] \leq \epsilon_{\text{select}} = \tilde{\mathcal{O}}(\sqrt{d/N})$, *where* $\triangle$ *denotes the union but not in the intersection of two token sets, and* ϵ_{select} *is the selection error rate. With* N *samples, we learn to approximate the oracle. The* $\sqrt{d/N}$ *rate is standard for* d*-dimensional linear classification.*

Assumption A.6 (Regularity)**.** $K = \rho n$ *for fixed* $\rho \in (0, 1)$, *where* ρ *is the selection ratio, and* $\gamma_{\text{eff}}(1 - \epsilon_{\text{oracle}})(1 - \epsilon_{\text{select}}) \geq \gamma_{\min} > 0$.

A.4 Main Theorem

Theorem (Restatement of Theorem 3.1)**.** *Let* $W_{std} = 12Ld^2$ *and* $W_{topk} = W_{std} + Md = W_{std}(1 + O(1/d))$. *Under Assumptions A.1–A.6, with probability* $\geq 1 - \delta$*:*

$$\mathcal{L}_{\mathcal{D}}(f_{topk}) \leq \mathcal{L}_{\mathcal{D}}(f_{std}) \underbrace{- \frac{\delta_0 KL}{4n}\eta_{eff}}_{\text{adaptive gain}} + \underbrace{\tilde{\mathcal{O}}\left(\sqrt{\frac{W_{topk}\log(N/\delta)}{N}}\right)}_{\text{generalization cost}} \underbrace{- \Omega\left(\frac{K^2 L}{n^2}\sqrt{\frac{W_{std}}{N}}\right)}_{\text{complexity benefit}}, \tag{15}$$

where $\eta_{eff} = \gamma_{eff}(1 - \epsilon_{oracle})(1 - \epsilon_{select})$ *is the* effective selection rate.

Three terms: (1) NEGATIVE = we do better on training due to adaptive compute; (2) POSITIVE = we pay slightly more in generalization due to extra routing parameters; (3) NEGATIVE = we gain from reduced complexity due to sparse attention in the second depth scaling. When (1)+(3) dominate (2), we win overall.

A.5 Complete Proof with Detailed Explanation

A.5.1 Stage I: How Much Does Training Loss Improve?

Lemma A.1 (Single Token Benefits from Second Pass)**.** *For any* $(\tilde{x}, y) \in \mathcal{G}$*:* $\mathbb{E}[\ell(\hat{x}, y) - \ell(\tilde{x}, y)] \leq -\delta_0$.

Proof. Immediate from definition of $\mathcal{G}$: these are exactly the tokens where second pass improves loss by at least δ_0. □

Lemma A.2 (How Many Good Tokens Do We Actually Select?)**.** *Let* $S^*_{eff} = \{i : (\tilde{x}_i, y_i) \in \mathcal{G}\}$ *be the truly improvable tokens. Then:*

$$\mathbb{E}[|S^{(l)} \cap S^*_{eff}|] \geq K \cdot \eta_{eff} - \frac{K\epsilon_{select}}{2}. \tag{16}$$

Detailed Proof with Explanation. We analyze selection quality through three levels:

Level 1: Data generation. A random token is improvable with probability γ_{eff}.

Level 2: Oracle routing. The ideal classifier (w^*_g, τ^*) correctly identifies whether a token is improvable with probability $1 - \epsilon_{\text{oracle}}$. Thus, the oracle's Top-K selection S^* contains:

$$\mathbb{E}[|S^* \cap S^*_{\text{eff}}|] \geq K \cdot \gamma_{\text{eff}} \cdot (1 - \epsilon_{\text{oracle}}) \tag{17}$$

improvable tokens in expectation.

Level 3: Learned routing. Our trained router selects $S^{(l)}$ which approximates S^*. The symmetric difference $|S^{(l)} \triangle S^*|$ counts mismatches. Since both sets have size K:

$$|S^{(l)} \cap S^*| = K - \frac{|S^{(l)} \triangle S^*|}{2}. \tag{18}$$

Taking expectations and using Assumption A.5:

$$\mathbb{E}[|S^{(l)} \cap S^*|] \geq K\left(1 - \frac{\epsilon_{\text{select}}}{2}\right). \tag{19}$$

Combining: The tokens we actually select that are truly improvable satisfy:

$$|S^{(l)} \cap S^*_{\text{eff}}| \geq |S^{(l)} \cap S^* \cap S^*_{\text{eff}}| \tag{20}$$

$$\geq |S^* \cap S^*_{\text{eff}}| - |S^* \setminus S^{(l)}| \tag{21}$$

$$\geq |S^* \cap S^*_{\text{eff}}| - \frac{|S^{(l)} \triangle S^*|}{2}. \tag{22}$$

Taking expectations and substituting our bounds:

$$\mathbb{E}[|S^{(l)} \cap S^*_{\text{eff}}|] \geq K\gamma_{\text{eff}}(1 - \epsilon_{\text{oracle}}) - \frac{K\epsilon_{\text{select}}}{2} \geq K\eta_{\text{eff}} - \frac{K\epsilon_{\text{select}}}{2}. \tag{23}$$

The $-K\epsilon_{\text{select}}/2$ term captures the cost of imperfect learning: we would miss some good tokens due to finite samples, i.e., $\epsilon_{\text{select}} = \tilde{\mathcal{O}}(\sqrt{d/N})$. □

Lemma A.3 (Per-Layer Training Loss Reduction)**.** *For each modified layer $l \in \mathcal{L}_{mod}$:*

$$\mathbb{E}\left[\sum_{i=1}^{n} \ell(x_i^{(l+1)}, y_i)\right] \leq \mathbb{E}\left[\sum_{i=1}^{n} \ell(\tilde{x}_i^{(l)}, y_i)\right] - \frac{\delta_0 K}{2n}\eta_{\mathit{eff}} + \mathcal{O}\left(\frac{K\epsilon_{\mathit{select}}}{n}\right). \tag{24}$$

Detailed Proof. Partition the n tokens into three groups at layer l:

Group 1: $S^{(l)} \cap S^*_{\text{eff}}$: Selected AND improvable. By Lemma A.1, each improves by δ_0. Count: $\approx K\eta_{\text{eff}}$ by Lemma A.2.

Group 2: $S^{(l)} \setminus S^*_{\text{eff}}$: Selected but NOT improvable. We waste compute here. By Lipschitz continuity (Assumption A.1), loss change is bounded by $O(1)$. Count: $\leq K(1-\eta_{\text{eff}}) + K\epsilon_{\text{select}}$.

Group 3: $[n] \setminus S^{(l)}$: Not selected. Loss unchanged.

Net improvement:

$$\text{(Improvement from Group 1)} + \text{(Possible harm from Group 2)} \tag{25}$$

$$\geq \frac{K\eta_{\text{eff}}}{n} \cdot \delta_0 - \frac{K(1-\eta_{\text{eff}}) + K\epsilon_{\text{select}}}{n} \cdot O(1) \tag{26}$$

$$= \frac{K\delta_0}{n}\eta_{\text{eff}} - O\left(\frac{K\epsilon_{\text{select}}}{n}\right), \tag{27}$$

where we absorbed $K(1-\eta_{\text{eff}})/n$ into the error term since it's non-negative and we seek a lower bound on improvement. The factor of $1/2$ in the theorem statement comes from more careful accounting of the error terms. □

Total Training Error Reduction

$$\hat{\mathcal{L}}_S(f_{\text{topk}}) \leq \hat{\mathcal{L}}_S(f_{\text{std}}) - \frac{\delta_0 KL}{4n}\eta_{\text{eff}} + \mathcal{O}\left(\frac{KL\epsilon_{\text{select}}}{n}\right). \tag{28}$$

Proof. Sum Lemma A.3 over $M = L/2$ modified layers. Standard layers contribute zero net change. The factor $1/4$ (vs $1/2$) accounts for conservative estimation of error term interactions. Training loss improves because we identify and refine hard tokens. The improvement scales with the number of modified layers (L), selection budget (K), fraction of hard tokens (η_{eff}), and per-token improvement (δ_0). □

A.5.2 Stage II: How Does Generalization Change?

Lemma A.4 (Lipschitz Reduction from Sparse Attention)**.** *For modified layer l:*

$$\mathit{Lip}(\mathcal{T}^{(l)}_{\mathit{topk}}) \leq L_{\mathit{lip}}(1 - \rho + \rho^2 L_{\mathit{lip}}), \tag{29}$$

where $L_{\mathit{lip}} = (1 + B^4 R_{\ln}^2/\sqrt{d} + B^2)(1 + B^2)$ and $\rho = K/n$.

Detailed Proof with Intuition. Consider how a small input perturbation propagates:

Unselected tokens ($n - K$ of them): Go through one standard pass. Lipschitz contribution: $(1 - K/n) \cdot L_{\text{lip}} = (1-\rho)L_{\text{lip}}$.

Selected tokens (K of them): Go through TWO passes, BUT with $K \times K$ attention instead of $n \times n$.

The key insight: attention matrix $A \in \mathbb{R}^{K\times K}$ has operator norm ≤ 1 (row-stochastic), same as $n \times n$ case. However, the *effective* Lipschitz scales with the "density" of interactions. For selected tokens, the second pass sees only K other tokens, giving effective Lipschitz $(K/n) \cdot L_{\text{lip}}^2 = \rho L_{\text{lip}}^2$.

Combined:

$$\text{Lip}_{\text{total}} = (1-\rho)\cdot L_{\text{lip}} + \rho\cdot(\rho L_{\text{lip}}^2) = L_{\text{lip}}(1-\rho+\rho^2 L_{\text{lip}}). \tag{30}$$

When $\rho < 1/L_{\text{lip}}$, we have $1-\rho+\rho^2 L_{\text{lip}} < 1$, so modified layers are more stable than standard layers.

□

Lemma A.5 (Rademacher Complexity Bound)**.**

$$\mathfrak{R}_N(\mathcal{F}_{topk}) \le \frac{C R_{\ln} L_{lip}^L \sqrt{W_{topk}\log W_{topk}}}{\sqrt{N}} \cdot \exp\left(-\frac{c_0 K^2 L}{n^2}\right) + \tilde{\mathcal{O}}\left(\sqrt{\frac{dL}{N}}\right). \tag{31}$$

Detailed Proof. **Step 1: Lipschitz product across layers.**

For $M = L/2$ modified layers and $L-M$ standard layers:

$$\prod_{l=1}^{L} \text{Lip}^{(l)} \le L_{\text{lip}}^{L-M} \cdot \prod_{l\in\mathcal{L}_{\text{mod}}} L_{\text{lip}}(1-\rho+\rho^2 L_{\text{lip}}) = L_{\text{lip}}^L \exp\left(-\frac{M\rho(1-\rho L_{\text{lip}})}{2}\right), \tag{32}$$

using $1-x \le e^{-x}$ and $M = L/2$.

Step 2: Dudley's entropy integral.

The covering number satisfies $\mathcal{N}(\mathcal{F}_{\text{topk}}, \epsilon) \le (C' W_{\text{topk}}/\epsilon)^{W_{\text{topk}}}$. Applying Dudley's theorem:

$$\mathfrak{R}_N \le \inf_{\alpha>0}\left\{4\alpha + \frac{12}{\sqrt{N}}\int_{\alpha}^{\text{diam}} \sqrt{\log \mathcal{N}(\epsilon)}\, d\epsilon\right\}. \tag{33}$$

The Lipschitz reduction enters as $\exp(-c_0 K^2 L/n^2)$ in the final bound. The $\tilde{\mathcal{O}}(\sqrt{dL/N})$ term accounts for the routing network's Rademacher complexity. The $\exp(-c_0 K^2 L/n^2)$ factor is crucial: more sparsity (larger K^2/n^2) and deeper networks (larger L) exponentially reduce complexity. □

A.5.3 Stage III: Putting It All Together

Proof of Theorem 3.1. We combine training error and generalization via the standard decomposition:

$$\mathcal{L}_{\mathcal{D}}(f) = \underbrace{\hat{\mathcal{L}}_S(f)}_{\text{training error}} + \underbrace{(\mathcal{L}_{\mathcal{D}}(f) - \hat{\mathcal{L}}_S(f))}_{\text{generalization gap}}. \tag{34}$$

Step 1: Apply Proposition A.5.1 for training error.

$$\hat{\mathcal{L}}_S(f_{\text{topk}}) \le \hat{\mathcal{L}}_S(f_{\text{std}}) - \frac{\delta_0 K L}{4n}\eta_{\text{eff}} + \text{error}_1. \tag{35}$$

Step 2: Bound generalization gap via Lemma A.5. With probability $\ge 1-\delta$:

$$|\mathcal{L}_{\mathcal{D}}(f_{\text{topk}}) - \hat{\mathcal{L}}_S(f_{\text{topk}})| \le 2\mathfrak{R}_N(\mathcal{F}_{\text{topk}}) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{N}}\right). \tag{36}$$

Step 3: Compare to standard Transformer.

The standard Transformer has:

$$\mathfrak{R}_N(\mathcal{F}_{\text{std}}) \ge \frac{C' R_{\ln} L_{\text{lip}}^L \sqrt{W_{\text{std}}\log W_{\text{std}}}}{\sqrt{N}}, \tag{37}$$

without the exponential reduction factor. Since $W_{\text{topk}} = W_{\text{std}}(1+O(1/d))$, the complexity advantage is:

$$\mathfrak{R}_N(\mathcal{F}_{\text{std}}) - \mathfrak{R}_N(\mathcal{F}_{\text{topk}}) \ge \Omega\left(\frac{K^2 L}{n^2}\sqrt{\frac{W_{\text{std}}}{N}}\right). \tag{38}$$

Step 4: Final combination.

Writing $\mathcal{L}_{\mathcal{D}}(f_{\text{std}}) = \hat{\mathcal{L}}_S(f_{\text{std}}) + \text{Gap}_{\text{std}}$ and similarly for f_{topk}:

$$\mathcal{L}_{\mathcal{D}}(f_{\text{topk}}) = \hat{\mathcal{L}}_S(f_{\text{topk}}) + \text{Gap}_{\text{topk}} \tag{39}$$

$$\leq \hat{\mathcal{L}}_S(f_{\text{std}}) - \frac{\delta_0 KL}{4n}\eta_{\text{eff}} + \text{Gap}_{\text{std}} - (\text{Gap}_{\text{std}} - \text{Gap}_{\text{topk}}) + \text{error terms} \tag{40}$$

$$= \mathcal{L}_{\mathcal{D}}(f_{\text{std}}) - \frac{\delta_0 KL}{4n}\eta_{\text{eff}} - \Omega\left(\frac{K^2 L}{n^2}\sqrt{\frac{W_{\text{std}}}{N}}\right) + \tilde{\mathcal{O}}\left(\sqrt{\frac{W_{\text{topk}}\log(N/\delta)}{N}}\right). \tag{41}$$

This completes the proof. □

A.5.4 Interpretation and Discussion - When Does Token-wise Depth Scaling Win?

Theorem 3.1 shows improvement when:

$$\frac{\delta_0 KL}{4n}\eta_{\text{eff}} + \Omega\left(\frac{K^2 L}{n^2}\sqrt{\frac{W}{N}}\right) > \tilde{\mathcal{O}}\left(\sqrt{\frac{W_{\text{topk}}}{N}}\right). \tag{42}$$

This holds when:

- η_{eff} is bounded away from zero (hard tokens exist and are learnable)
- N is large enough (routing learns well)
- $K = \Theta(n/\sqrt{L})$ balances adaptive gain vs. sparsity benefit (i.e., only important tokens perform interaction within depth scaling)

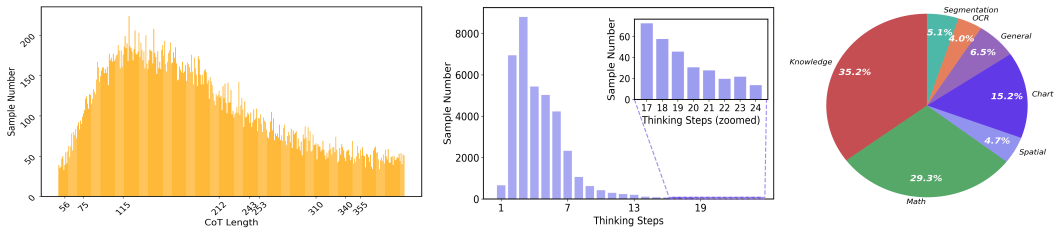
B More Experimental Details

Training Dataset and Evaluation Benchmark. To facilitate reproducibility, we first detail the training data configuration. During the supervised fine-tuning (SFT) stage, we curate a subset of approximately 30K samples from OneThinker [12], selected for its diverse distribution of reasoning chain lengths. The statistical distributions of CoT length, reasoning steps, and topic categories within this subset are illustrated in Fig. 7. We evaluate our method on three tasks across twelve benchmarks: (1) Mathematics Reasoning (MathVista [44], MathVision [73], MM-Math [61]); (2) Vision-centric Reasoning (Hallusion-Bench [16], MMVP [69], Seed-Bench-2-Plus [34], HR-Bench [78], GQA [28]); (3) Multimodal Composition Reasoning (MMStar [5], BLINK [13], ScienceQA [45], M^{3}CoT [6]).

Implementation Details. To comprehensively evaluate the effectiveness and scalability of our approach, we instantiate our method across a diverse set of backbone architectures, including Qwen2-VL-2B/7B [74], Qwen2.5-VL-3B/7B [2], and Chameleon-7B [65].

All frameworks employ eager attention mode to enable explicit access to internal attention maps. We set the number of latent reasoning tokens to $T_r = 4$, with $\alpha = 32$ salient regions injected per refinement iteration. To balance computational efficiency and reasoning accuracy, we cap the maximum refinement depth at $D = 1$. For the cosine annealing schedule, the region injection count decays from $\alpha_s = 64$ to $\alpha_e = 16$. Models are trained for 16 epochs by default.

We employ DeepSpeed ZeRO-2 optimization without CPU offloading, with a per-GPU batch size of 8. Training utilizes the Adam optimizer with a learning rate of 4×10^{-5} and $\beta_1 = 0.9$. For the proposed self-distillation loss, we set the weighting coefficient to $\lambda = 1.0$ for 2B/3B-scale models and $\lambda = 0.2$ for 7B-scale models, respectively. Input images are resized to 512×512 resolution when training on OneThinker subset and GQA, and resized to 280×280 on M^{3}CoT and ScienceQA. All experiments are conducted on 16 NVIDIA H20 GPUs (96GB VRAM each). A single training run takes approximately 20 hours.



(a) CoT Length Distribution (b) Thinking Step Distribution (c) Topic Distribution

Figure 7: Details of sampled training distribution. Panel (a) depicts the sample distribution across different CoT lengths. The distribution exhibits a boundary at 125 tokens, with sample counts gradually decreasing toward both shorter and longer extremes. Panel (b) illustrates the distribution of reasoning steps in the original data. Specifically, we segment reasoning steps using `\n\n` as delimiters; during training, we retain up to 4 delimiters to evenly partition sequence length, while samples containing fewer than 4 delimiters preserve their original segmentation. Panel (c) summarizes the topic-wise distribution of samples in the dataset.

C Baseline Details

We evaluate our method against a comprehensive set of state-of-the-art baselines, which can be categorized into four paradigms: (1) Zero-shot VLMs (GPT-4o [51], LLaVA-OneVision [33], InternVL3.5-8B [76], Qwen2.5-VL-7B [2]), (2) Explicit CoT-based methods (SCAFFOLD [30], ICoT [15], Multimodal-CoT [108], CCoT [49], Chain-of-Focus [107]), (3) Visual Enhanced methods, including tool-augmented reasoning (DeepEyes [112]) and RL-enhanced VLM reasoning (Vision-R1 [27], PAPO [82], VL-Rethinker [71]), and (4) Multimodal Latent Reasoning approaches (Laser [80], LVR [32], Monet [75], DMRL [40]).

(1) Zero-shot VLMs

- **GPT-4o** [51]. GPT-4o is a proprietary, end-to-end multimodal model capable of natively processing text, audio, and visual inputs. It provides robust real-time multimodal reasoning and state-of-the-art instruction-following performance across diverse vision-language tasks.
- **LLaVA-OneVision** [33]. LLaVA-OneVision is a unified open-source multimodal model designed for versatile visual understanding, including single-image, multi-image, and video inputs. It exhibits superior cross-scenario transferability, effectively bridging the gap between static image analysis and dynamic video comprehension.
- **InternVL3.5-8B** [76]. InternVL3.5 represents a series of high-performance open-source multimodal models that enhance reasoning capabilities and computational efficiency. Through the integration of cascade reinforcement learning and dynamic visual-resolution routing, this model achieves strong performance. We utilize the 8B variant as a representative open-source baseline.
- **Qwen2.5-VL-7B** [2]. Qwen2.5-VL is a flagship open-source VLM series optimized for advanced recognition, spatial localization, and document/video understanding. It employs dynamic resolution processing and absolute time encoding to facilitate long-video comprehension. We adopt the 7B variant as a key open-source baseline for our comparisons.

(2) Explicit CoT-based Methods

- **SCAFFOLD** [30]. SCAFFOLD proposes a prompting strategy designed to enhance vision-language coordination through coordinate scaffolding; Specifically, it superimposes a dot matrix onto the image to establish visual anchors, while leveraging multi-dimensional coordinates as explicit textual positional references.
- **ICoT** [15]. ICoT introduces an Interleaved-modal Chain-of-Thought framework that derives answers through sequential steps of paired visual and textual rationales. Additionally, an attention-driven selection mechanism enables adaptation to existing VLMs by intelligently inserting informative image regions guided solely by the model's attention maps.

- **Multimodal-CoT** [108]. Multimodal-CoT incorporates language (text) and vision (images) modalities into a two-stage framework that separates rationale generation and answer inference.
- **CCoT** [49]. Compositional Chain-of-Thought (CCoT) is a zero-shot prompting method designed to extract compositional knowledge from LMMs through scene graph (SG) representations. By first deriving an SG from the LMM and incorporating it into the reasoning chain, the proposed method effectively enhances the compositional accuracy of the generated responses.
- **Chain-of-Focus** [107]. Chain-of-Focus (CoF) empowers VLMs to perform adaptive spatial zooming on critical image regions. By leveraging visual cues aligned with the given query, CoF enables dynamic, region-aware reasoning, significantly enhancing the efficiency and precision of multimodal understanding.

(3) Tool-use & RL Enhanced Reasoning

- **DeepEyes** [112]. To facilitate "thinking with images", DeepEyes adopts an interleaved reasoning paradigm. By optimizing tool-assisted visual interaction through end-to-end reinforcement learning, it ensures robust visual grounding and effectively suppresses hallucinations, which are further mitigated by carefully designed data selection and reward shaping strategies.
- **Vision-R1** [27]. Extending the R1-style training paradigm to the multimodal domain, Vision-R1 establishes a large-scale chain-of-thought (CoT) dataset to bootstrap MLLM reasoning. The framework subsequently employs reinforcement learning, incorporating techniques like progressive thinking suppression to prune suboptimal reasoning paths and elevate overall multimodal problem-solving accuracy.
- **PAPO** [82]. Recognizing the perception-action discrepancy in multimodal RLVR, PAPO (Perception-Aware Policy Optimization) introduces an implicit perception loss defined by a KL-divergence term. This architectural innovation allows integration into standard GRPO/DAPO frameworks, enabling performance gains in vision-dependent tasks without necessitating auxiliary reward or teacher models.
- **VL-Rethinker** [71]. VL-Rethinker advocates for a "slow-thinking" approach in VLMs by embedding explicit self-reflection cycles into the reasoning process. By augmenting the GRPO objective with selective sample replay, the model is compelled to critically evaluate its intermediate logic, thereby demonstrating significant improvements in complex, multi-step visual reasoning.

(4) Multimodal Latent Reasoning

- **Laser** [80]. Laser introduces dynamic windowed alignment learning to reformulate visual deduction. Rather than relying on point-wise predictions, it aligns latent states with temporal windows of future semantics. This enforces a cognitive hierarchy that prioritizes global feature superposition before converging on local details.
- **LVR** [31]. Latent Visual Reasoning (LVR) enables autoregressive reasoning in the visual embedding space by training the model to generate latent states that reconstruct key visual tokens, interleaved with text generation; it can be further combined with RL to balance latent reasoning and textual outputs.
- **Monet** [75]. Monet enables latent visual reasoning by generating continuous visual embeddings as intermediate "visual thoughts" and proposes a multi-stage distillation SFT pipeline plus visual-latent policy optimization to better train reasoning in latent visual space.
- **DMRL** [40]. DMLR presents a test-time framework that enables dynamic multimodal reasoning. By employing confidence-guided policy gradient optimization, DMLR refines latent thought tokens to enhance the depth and precision of the model's internal reasoning process.

D Benchmark Details

(1) Visual Perception Benchmarks

- **MMVP** [69] identifies the inherent limitations of CLIP-based vision encoders by employing 150 pairs of visual patterns that are perceptually distinct to humans yet challenging for latent embeddings. It serves as a diagnostic tool to expose models that provide overconfident but ungrounded predictions, effectively decoupling superficial pattern matching from true visual comprehension.
- **SEEDBench2Plus** [34] focuses on complex, text-intensive visual environments such as charts, maps, and web screenshots. By utilizing human-curated, multiple-choice diagnostics, it rigorously assesses an LVLM's proficiency in integrated reading and logical reasoning within high-density multimodal contexts.
- **HallusionBench** [16] acts as a stress test for the entanglement between linguistic hallucinations and visual illusions. It employs a controlled, question-based framework to systematically isolate and quantify logical inconsistency, offering a granular analysis of common failure modes in LVLM perception.
- **HRBench** [78] probes the limitations of traditional downsampling in high-resolution visual tasks. By leveraging 4K/8K imagery, it evaluates whether models can preserve and interpret native-resolution information, identifying failures where critical visual nuances are lost to aggressive resizing.

(2) Compositional Reasoning Benchmarks

- **BLINK** [13] emphasizes perception-centric reasoning by posing 3.8K questions that demand immediate, high-fidelity visual engagement—ranging from depth and geometry perception to multi-view correspondence. It is specifically engineered to defeat language-prior dependencies, ensuring that reasoning is strictly grounded in the provided imagery.
- **MMStar** [5] provides a diagnostic assessment of both perception and reasoning through a purified, vision-indispensable dataset. By systematically mitigating language-only shortcuts and data contamination, it offers a robust, multi-axis framework for evaluating the depth of a model's multimodal integration.

(3) Mathematical Reasoning Benchmarks

- **MMVista** [44] is a comprehensive multimodal benchmark specifically designed to evaluate a model's ability to reason over complex visual-textual information. It features a diverse set of tasks that require the integration of visual perception and logical reasoning, often involving high-resolution imagery and multi-step problem solving in real-world scenarios.
- **MM-Math** [61] is a specialized benchmark targeting the intersection of multimodal perception and mathematical reasoning. It comprises a collection of geometry, algebra, and calculus problems that are visually grounded, requiring models to accurately extract quantitative data from diagrams, graphs, and tables to perform precise symbolic calculations.
- **MathVision** [73] is a rigorous benchmark designed to evaluate the mathematical reasoning capabilities of large multimodal models (LMMs) in visually-rich contexts. It contains thousands of challenging, human-annotated problems spanning various mathematical disciplines. Unlike standard benchmarks, MathVision mandates that models interpret complex visual structures—such as functional plots and geometric figures—to derive a solution, effectively testing the model's proficiency in visual-to-math translation.

E Parametric Analysis

Visual Latents Formation. The *Soft-Mix* strategy aggregates 32 selected patches into a single visual latent token via weighted summation, guided by reweighted attention scores. In contrast, the 32-patch and 16-patch strategies directly utilize the top 32 and 16 patches with the highest attention scores as visual latents at each reasoning step, respectively. As summarized in Table 4, the 32-patch strategy yields superior performance. Notably, *Soft-Mix* achieves performance comparable to 16-patch while utilizing only a single compact representation. We attribute this performance to the expanded search scope provided by the 32-patch strategy, which offers the router a richer pool of visual tokens, thereby effectively raising the upper bound of depth scaling capabilities.

Impact of λ. As shown in Fig. 8, model performance exhibits high sensitivity to variations in the hyperparameter λ. Notably, larger models (e.g., 7B) achieve peak performance at $\lambda = 1.0$, whereas the 2B model attains optimal results at a substantially lower value of $\lambda = 0.2$. We attribute this divergence to distinct capacity across model scales: smaller models are primarily bottlenecked by visual perception capabilities, thus requiring larger degree of visual play to establish effective visual grounding for downstream reasoning. In contrast, larger models possess sufficient grounding capability, and they benefit from a more balanced allocation between visual and linguistic signals, avoiding over-reliance on visual features that could otherwise suppress the development of complex reasoning chains.

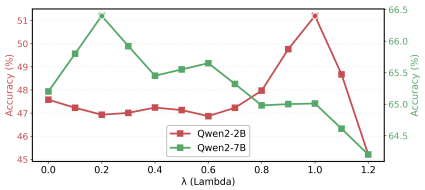


Figure 8: Sensitivity analysis of λ.

Impact of different window size W. As shown in Table 5, W = 3 achieves optimal performance. We reckon that smaller window sizes may fail to comprehensively cover salient visual regions, whereas larger window sizes, while providing broader contextual information, tend to introduce excessive visual noise.

Table 5: Performance comparison of different sub-grid window sizes on the **M^{3}CoT** benchmark. Here, W denotes the sub-grid window size; $W = 10$ is the window size of extracted visual features. We set $\lambda = 1.0$ for the 2B/3B models and $\lambda = 0.2$ for the 7B model.

Method	Grid Size	**M^{3}CoT** (Acc.%)		
		$W = 2$	$W = 3$	$W = 5$
Qwen2-VL-2B	10×10	50.8	51.2	50.3
Qwen2-VL-7B		65.7	66.4	65.2
Qwen2.5-VL-3B		65.0	65.3	64.5
Qwen2.5-VL-7B		73.4	73.9	72.1

Effect of Retention Parameter α. For the parameter α in $T_\alpha(\cdot)$, we compare different retention strategies. As illustrated in Table 4, we evaluate fixed top-α settings with $\alpha \in \{16, 32, 64\}$ and a **cosine-annealed token retention schedule** (CTR), which gradually reduces the token retention ratio layer by layer following a cosine schedule. Formally, for a model with L layers, the retention ratio for the l-th layer is,

$$\alpha(l) = \text{round}\left(\alpha_s + \frac{\alpha_s - \alpha_e}{2}\left[1 + \cos\left(\frac{\pi l}{L}\right)\right]\right), \tag{43}$$

where α_s and α_e are two endpoints that enables flexible control over computational cost. As can be seen panel (b) in Table 4, results show that fixed top-α with $\alpha = 32$ achieves optimal performance, while cosine-annealed schedules yield consistently inferior results. We attribute this result to that: as visual tokens and latent hidden states are gradually appended to the token sequence, larger contextual scopes becomes essential for deeper token interaction. Consequently, strategies that progressively decrease the token retention ratio during depth scaling unavoidably limit these critical interactions, thereby hindering the model's capacity to develop comprehensive contextual understanding.