Li Auto | Base Model

Self-Consistent Latent Reasoning: Long Latent Sequence Reasoning for Vision-Language Model

Chenfeng Wang[1,2] **Wei He**[2] **Xuhan Zhu**[2] **Chunpeng Zhou**[2] **Qizhen Li**[2]
Song Yan[1] **Yufei Zheng**[1,2] **Chengjun Yu**[1,2] **Fan Lu**[1]
Wei Zhai[1,†] **Yang Cao**[1]
Pengfei Yu[2,†] **Zheng-Jun Zha**[1]

[1]University of Science and Technology of China [2]Li Auto Inc.

In language reasoning, longer chains of thought consistently yield better performance, which naturally suggests that visual latent reasoning may likewise benefit from longer latent sequences. However, we discover a counterintuitive phenomenon: the performance of existing latent visual reasoning methods systematically *degrades* as the latent sequence grows longer. We reveal the root cause: **Information Gain Collapse**—autoregressive generation makes each step highly dependent on prior outputs, so subsequent tokens can barely introduce new information. We further identify that heavily pooled ($\geq$128$\times$) image embeddings used as supervision targets provide no more signal than meaningless placeholders. Motivated by these insights, we propose **SCOLAR** (**S**elf-**CO**nsistent **LA**tent **R**easoning), which introduces a lightweight *detransformer* that leverages the LLM's full-sequence hidden states to generate auxiliary visual tokens *in a single shot*, with each token independently anchored to the original visual space. Combined with three-stage SFT and **ALPO** reinforcement learning, SCOLAR extends acceptable latent CoT length by over 30$\times$, achieves state-of-the-art among open-source models on real-world reasoning benchmarks (+14.12% over backbone), and demonstrates strong out-of-distribution generalization.

Date: May 13, 2026
Correspondence: yupengfei1@lixiang.com, wzhai056@ustc.edu.cn
Project Page: https://github.com/SCOLAR866/SCOLAR

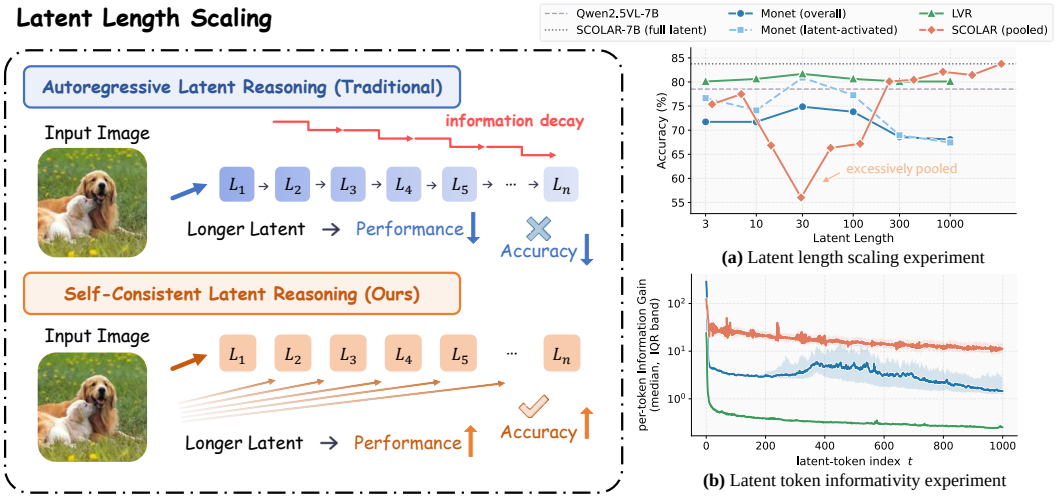
1 Introduction

In recent years, the "compute-for-performance" scaling philosophy has fundamentally reshaped the paradigm of language reasoning: longer **CoT** (chains of thought) and more reasoning steps have led to consistent capability leaps in language models [4, 13, 20, 55]. This principle naturally prompts researchers to ask: *can visual reasoning similarly benefit from longer latent reasoning chains?* Recent work [14, 23, 45, 54] on latent reasoning in **VLMs** (vision-language models) [1, 5, 7, 9, 21, 37, 47] explores precisely this direction—introducing visual reasoning steps in the form of latent representations during generation, with the expectation that richer accumulated visual semantics will yield stronger reasoning capability.

However, we discover a puzzling, counterintuitive phenomenon. As shown in Figure 1(a), the performance of existing latent visual reasoning methods [23, 45] systematically degrades as the latent sequence grows longer—a longer reasoning chain is a burden, not a benefit. This stands in sharp

†Corresponding author.

Preprint.



(a) Latent length scaling experiment

(b) Latent token informativity experiment

Figure 1: **Latent Length Scaling: Phenomenon, Root Cause, and Solution. Left.** Comparison of conventional autoregressive latent reasoning and self-consistent latent reasoning: In conventional methods, the generation of the latent variable L_n depends solely on the preceding latent L_{n-1}, so effective semantic information in latent tokens gradually decays along the chain (**information decay**), causing longer sequences to degrade performance. In SCOLAR, each auxiliary token independently anchors to the original visual semantic space, fundamentally preventing information decay, and longer sequences continuously improve performance. **Right (a).** Latent length scaling experiments: SCOLAR continuously improves beyond a certain latent length threshold, while Monet and LVR systematically degrade. **Right (b).** Latent token information gain experiments: SCOLAR maintains a consistently high information gain over the 1K-token sequence, while LVR rapidly decays to near zero after $t \approx 10$, and Monet exhibits early saturation.

contrast to the intuition from language CoT scaling: we expect "more compute = better reasoning," yet in the visual latent space, "longer chain = worse performance."

What is the root cause of this phenomenon? We systematically diagnose it from multiple dimensions through three complementary experiments. *(1) Length scaling experiments* (Figure 1(a)): We measure accuracy curves as latent length varies from 3 to 1000+, quantifying the degradation pattern. *(2) Meaningless padding experiments*: Replacing the latent tokens generated by LVR [23] with meaningless padding characters produces almost no performance change—if such replacement is lossless, the original tokens carry no real visual semantics. *(3) Information gain analysis* (Figure 1(b)): We quantify the incremental information contributed by each new latent token using orthogonal projection residuals [41], tracking information accumulation curves across the entire sequence.

All three experiments converge on the same root cause: **Information Gain Collapse**. Figure 1(b) clearly shows that LVR's information gain drops to near zero after sequence position $t \approx 10$, while Monet's [45] quickly saturates—neither can continuously accumulate new visual semantics as the chain grows. The fundamental mechanism is: *autoregressive generation* makes each prediction step highly dependent on its own prior outputs; as shown in the upper-left figure, original visual information inevitably decays as it propagates along the chain (information decay), and subsequent tokens can only repeat existing information. Additionally, our experiments reveal another overlooked deficiency: using heavily pooled ($\geq$128$\times$) image embeddings as latent supervision targets produces semantics equivalent to meaningless noise, which cannot effectively guide visual reasoning—directly challenging prior methods that rely on excessively pooled features as alignment targets.

The above analysis yields a clear design principle: to avoid information gain collapse, each auxiliary latent token must *independently anchor to the original visual semantic space*, rather than relying on transmission from other latent tokens in the sequence. Based on this insight, we propose **SCOLAR** (Self-COnsistent LAtent Reasoning): on the standard ViT–Projector–LLM framework [24, 26], we introduce a lightweight *detransformer* branch that directly leverages the LLM's *full-sequence hidden states* to generate, *in a single shot*, an auxiliary visual latent sequence fully aligned with the input resolution. As shown in the lower-left of Figure 1, each auxiliary token L_t independently connects back to the original image semantic space, completely breaking the autoregressive dependency chain.

Self-consistent means the auxiliary visual tokens remain anchored to the original visual semantic space throughout generation and use: complete visual features at the same resolution as the input serve as the alignment target, ensuring each auxiliary token carries genuine visual semantics; three-stage SFT (with teacher-forcing annealing) and **ALPO** reinforcement learning further alleviate implicit distribution mismatch, maintaining training–inference consistency in long-latent settings.

The effect of this design is direct: in Figure 1(a), SCOLAR's accuracy recovers from the moderate-pooling dip and ultimately peaks at full latent length; in Figure 1(b), SCOLAR maintains high information gain (approximately 10–30) throughout the entire 1000-token sequence, surpassing Monet and LVR by over an order of magnitude. SCOLAR extends the *acceptable maximum latent CoT length* by over 30×, achieves state-of-the-art among current open-source models on real-world reasoning benchmarks including MME-RealWorld-Lite [61] (+14.12% over the backbone), HRBench4K [48], and V*Bench [51], demonstrates strong out-of-distribution generalization on the visual logic reasoning benchmark VisualPuzzles [39], and surpasses GPT-4o [19] (67.50%) by over 16% on V*Bench with 83.77%. Our main contributions are:

- We identify *Information Gain Collapse* as the root cause of performance degradation in existing latent reasoning methods, and propose SCOLAR—a single-shot paradigm where a lightweight detransformer generates auxiliary visual tokens independently anchored to the original visual space, making latent length truly scalable.
- We design three-stage SFT with teacher-forcing annealing and ALPO reinforcement learning to resolve the distribution mismatch inherent in long-latent training.
- SCOLAR extends acceptable latent CoT length by over 30×, achieves state-of-the-art among open-source models (+14.12% on MME-RealWorld-Lite, surpassing GPT-4o by 16% on V* Bench), with strong OOD generalization.

2 Related Work

Think with Images. Existing multimodal reasoning methods enrich the reasoning process by introducing additional visual evidence, but they do so in two different ways. Some methods [22, 30, 46, 49, 52, 56] still reason mainly in *text* space, using SFT or RL to elicit longer text CoT while relying on images as supporting evidence. Others explicitly inject visual operations into the reasoning chain, for example by grounding, cropping [3, 18, 43, 59], re-inserting selected image tokens, or invoking external tools to draw, edit [6, 11, 17, 32, 34, 40, 60, 62, 64], or generate intermediate visual content [8, 16, 25, 53]. These approaches improve task-relevant perception, but they either remain fundamentally text-centric or depend on tool availability, invocation supervision, and often asynchronous multi-turn inference, which increases system complexity and latency.

Latent Space Reasoning. Another line of work moves reasoning from discrete tokens to continuous latent space. In NLP, some methods [2, 12, 14, 15, 44, 50] replace text tokens with continuous hidden states as intermediate reasoning steps. In MLLMs, some methods [23, 45, 54] align generated latent representations with auxiliary image embeddings, while others [31] propose to remove auxiliary images and constrain latent generation purely through next-token prediction. However, existing latent reasoning methods still share several limitations: they rely on autoregressive token-by-token generation, so information gain decays rapidly as the chain grows; some methods use excessively pooled image embeddings as supervision targets, which our experiments show can degrade into semantically weak or even meaningless signals. These limitations motivate our single-shot latent reasoning framework.

3 Method

3.1 Overview

Built on the standard ViT–Projector–LLM architecture, SCOLAR introduces a lightweight *detransformer* branch that generates **auxiliary visual tokens** aligned with the input resolution from the LLM's *full-sequence hidden states* in a single forward pass, and fuses them with original visual features to serve subsequent language reasoning. This **single-shot, globally consistent** latent generation is the key to scaling latent length without performance degradation. The overall pipeline is illustrated in Figure 2:

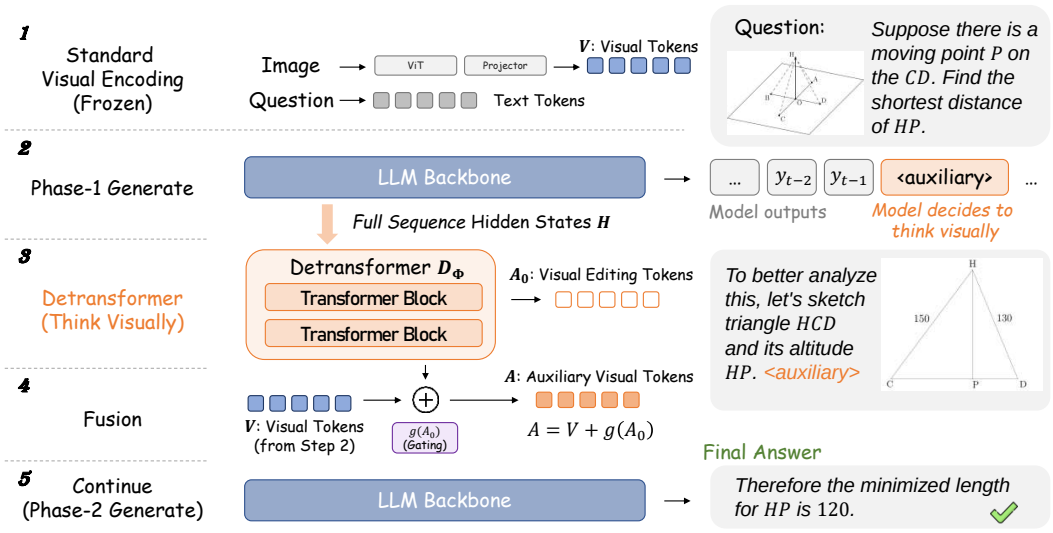


Figure 2: **Overview of the SCOLAR Inference Pipeline.** The model encodes visual tokens, generates Phase-1 text until the <auxiliary> trigger, produces auxiliary visual tokens via the detransformer in a single shot, fuses them with original visual features, and continues Phase-2 generation in the updated context.

Specifically, the inference proceeds in five steps: ① The input image and question are encoded into visual tokens V via ViT–Projector; ② Text tokens and V are fed into the LLM for Phase-1 generation until the model autonomously outputs the trigger token <auxiliary> (whether to trigger is entirely determined by the model's policy and may occur zero or multiple times); ③ The current full-sequence hidden states H are extracted and fed into the detransformer D_ϕ, generating auxiliary visual tokens A_0 in a single shot; ④ After gating, A_0 is fused with the original visual tokens: $A = V + g(A_0)$; ⑤ The LLM continues Phase-2 generation in the updated visual context, producing the final answer.

Formally, let $V = E_{\text{vis}}(I) \in \mathbb{R}^{N_v \times d_v}$, $X_v = P(V) \in \mathbb{R}^{N_v \times d}$, where I is the input image, V is the visual feature sequence, X_v is the projected visual token sequence, N_v is the number of visual tokens, and d is the LLM hidden dimension. Suppose the model has generated a prefix $y_{1:t}$ and emits the trigger token <auxiliary>. We then collect the full-sequence hidden states from a selected LLM layer, and the detransformer maps H to auxiliary visual tokens:

$$H = f_\theta^{(\ell)}([X_v; q; y_{1:t}; \texttt{<auxiliary>}]) \in \mathbb{R}^{L \times d}, \qquad A_0 = D_\phi(H) \in \mathbb{R}^{N_a \times d_v}, \tag{1}$$

where q is the question, L is the full sequence length, ℓ denotes the selected hidden layer, and N_a is the auxiliary latent length. In practice, $N_a = N_v$, i.e., the detransformer produces auxiliary tokens at the same count as the input visual tokens.

Detransformer Architecture The detransformer is implemented as a two-layer Transformer [42] that operates directly on the full-sequence hidden states H. It applies layer normalization, multi-head self-attention, and feed-forward stacking to H, producing auxiliary visual editing tokens A_0 at the same spatial resolution as the input visual tokens. The output is fused with original visual features through gated residual addition:

$$A = V + g(A_0), \qquad g(A_0) = \alpha\, A_0, \quad \alpha \in [0, 1], \tag{2}$$

where α controls the editing strength. Full architectural details are provided in Appendix A. After fusion, the LLM resumes decoding in the updated visual context:

$$y_{t+1} = \text{decode}\big(f_\theta([X_v; q; y_{1:t}; A])\big), \tag{3}$$

where $\text{decode}(\cdot)$ denotes passing hidden states through the Language Model Head to obtain logits and sampling the next token.

3.2 Supervised Fine-Tuning (SFT)

Training data consists of geometrically-aligned image–auxiliary-image–answer triples $(I, x, y^{(1)}, I_{\text{aux}}, y^{(2)})$, constructed via SIFT(Scale-invariant Feature Transform) [28, 29] alignment,

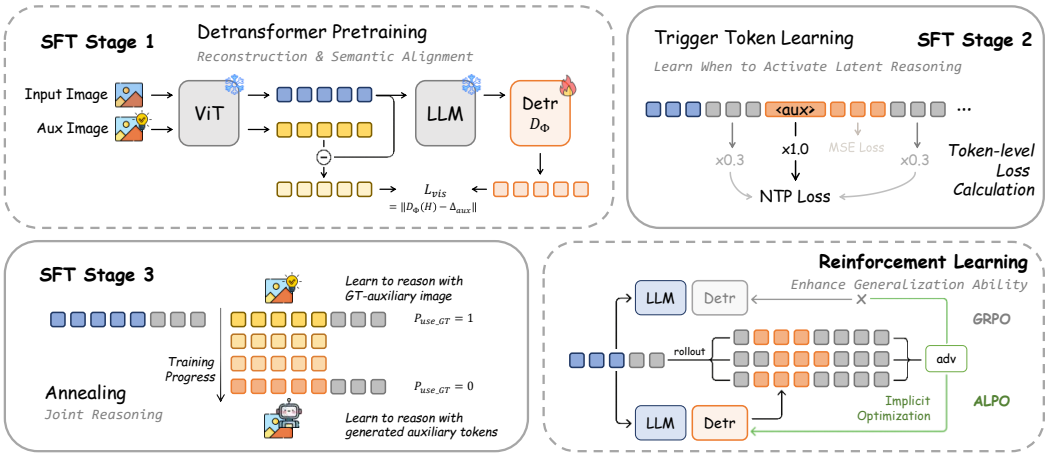


Figure 3: **Overview of the SCOLAR Training Pipeline.** Three SFT stages plus one RL stage. **Stage 1:** Detransformer pretraining with delta-feature reconstruction. **Stage 2:** Trigger token learning via weighted NTP. **Stage 3:** Joint reasoning with teacher-forcing annealing. **RL:** ALPO with two-phase rollout and outcome-driven rewards.

CoT distillation, and necessity filtering (details in Appendix B). Direct end-to-end training suffers from *implicit distribution mismatch*: ground-truth auxiliary visuals are used during training, but inference relies on *self-generated* latents. We therefore adopt **three-stage SFT**: (i) pre-training the detransformer for reconstruction, (ii) weighted NTP for trigger learning, and (iii) teacher-forcing annealing for smooth transition to self-generated auxiliary visuals. Figure 3 provides an overview.

Stage 1: Detransformer Pretraining. Only D_ϕ is trained; f_θ is frozen. To avoid an untrained visual editing branch directly coupling with the LLM and collapsing its semantic representations, we first constrain D_ϕ's output to align with the target auxiliary visual features using a **reconstruction loss**:

$$\Delta_{aux} = E_v(I_{\text{aux}}) - E_v(I), \qquad \mathcal{L}_{\text{reconst}} = \|D_\phi(H) - \Delta_{aux}\|_2^2. \tag{4}$$

where E_v is the Vision Encoder that converts input images into visual tokens. Here we use the "**Visual editing feature**" Δ_{aux} as the target. During training, inputs include the image I, question x, and text $y^{(1)}$ preceding `<auxiliary>`, to ensure D_ϕ is modulated by textual context rather than performing unconstrained arbitrary visual editing.

Stage 2: Learning the Trigger Token. In this stage, we train trigger behavior with **weighted next-token prediction** (NTP):

$$\mathcal{L}_{\text{ntp}} = -\sum_t w_t \log \pi_\theta(y_t \mid y_{<t}, V, x_{1:T}), \quad w_t = \begin{cases} w_{\text{aux}}, & y_t = \texttt{<auxiliary>} \\ w_{\text{normal}}, & \text{otherwise} \end{cases} \tag{5}$$

where $w_{\text{normal}} \le 1 \le w_{\text{aux}}$, imposing higher loss weight on the trigger token to prevent the tradeoff between learning rate and general representations from making triggers hard to learn.

Stage 3: Joint Reasoning with Auxiliary Visuals. In this stage, the model must, after generating `<auxiliary>`, continue to complete the answer *based on its self-generated* $A = D_\phi(H)$. We jointly optimize both language and visual objectives:

$$\mathcal{L} = \mathcal{L}_{\text{ntp}} + \lambda \mathcal{L}_{\text{vis}}. \tag{6}$$

To alleviate the distribution mismatch of "training with ground-truth auxiliary, testing with model-generated auxiliary," we apply **teacher forcing annealing**: at the start of training, the ground truth U is used as a substitute for A; subsequently, the probability of using the ground truth is linearly decayed from $1 \to 0$, gradually adapting the model to its own generated auxiliary visual features.

3.3 ALPO: Auxiliary Latents Policy Optimization

While three-stage SFT equips the model with basic latent reasoning capabilities, its distribution coverage is limited. We propose **ALPO** (Auxiliary Latents Policy Optimization) to further optimize

Table 1: **Performance on real-world perception and reasoning benchmarks.** The best-performing open-source model for each dataset is highlighted in **bold**, while those that performed second-best are marked with an underline.

Model	V*			HRBench4K			HRBench8K			MME-RealWorld-Lite		
	Overall	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall	Reasoning	Perception
GPT-4o [19]	67.5*	72.2*	60.5*	59.0*	70.0*	48.0*	55.5*	62.0*	49.0*	52.0*	48.3*	54.4*
GPT-5.2 [38]	63.87	61.74	67.11	66.37	71.00	61.75	61.50	64.75	58.25	50.23	45.87	53.04
Qwen2.5-VL-7B [35]	78.53	80.00	76.31	71.50	83.00	60.00	63.75	73.75	53.75	45.75	39.73	49.62
Deepeyes [63]	83.25	84.35	81.58	71.25	83.75	58.75	65.13	77.00	53.25	54.28	50.53	56.63
LVR [23]	80.6*	81.7*	79.0*	-	-	-	-	-	-	-	-	-
Monet [45]	83.25*	83.48*	**82.89***	71.00*	**85.25***	56.75*	**68.00***	**79.75***	56.25*	55.50*	51.07*	58.34*
SCOLAR(SFT only)	**84.29**	**86.09**	81.58	72.50	83.75	61.25	65.25	74.25	56.25	53.62	48	57.23
SCOLAR-7B	83.77	**86.09**	80.26	**75.50**	82.50	**68.50**	67.63	73.50	**61.75**	**59.87**	**55.14**	**63.13**
Δ (vs Qwen2.5VL-7B)	+5.24	+6.09	+3.95	+4.00	-0.50	+8.50	+3.88	-0.25	+8.00	+14.12	+15.41	+13.51

SCOLAR with outcome-driven reward signals. The key challenge is that SCOLAR inserts *continuous* auxiliary visual embeddings $e_{\text{aux}} \in \mathbb{R}^{N_{\text{vis}} \times d}$ between Phase 1 and Phase 2 generation, which have no discrete token distribution and cannot be incorporated into standard policy gradient frameworks (e.g., GRPO [13, 36]). ALPO addresses this by *decoupling* the continuous visual reasoning from the discrete text policy gradient: during rollout, e_{aux} is generated following the inference pipeline and **cached**; during the actor update, the cached e_{aux} is injected as **fixed context**, and the policy gradient is computed only over discrete tokens $o_1 \oplus o_2$:

$$\mathcal{J}_{\text{ALPO}}(\theta) = \mathbb{E}\left[\sum_t \min\Big(\tilde{r}_t(\theta)\hat{A}_t,\ \text{clip}(\tilde{r}_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\right], \tag{7}$$

$$\tilde{r}_t(\theta) = \frac{\pi_\theta(o_t \mid o_{<t},\ x, I,\ e_{\text{aux}})}{\pi_{\theta_{\text{old}}}(o_t \mid o_{<t},\ x, I,\ e_{\text{aux}})}, \tag{8}$$

where $\hat{A}_t$ is estimated via GRPO's within-group normalization over G sampled responses.

Reward Design. ALPO adopts a three-component reward:

$$r = r_{\text{acc}} + \lambda_{\text{fmt}} \cdot r_{\text{fmt}} + \lambda_{\text{aux}} \cdot r_{\text{aux}}, \tag{9}$$

where r_{acc} is the accuracy reward, r_{fmt} checks format correctness, and r_{aux} gives a positive reward if and only if the response correctly generates `<auxiliary>`. Since policy gradient cannot directly optimize continuous auxiliary features, r_{aux} implicitly encourages the model to activate the latent visual reasoning path by rewarding the discrete trigger token (see Appendix C for further discussion).

4 Experiments

4.1 Experimental Settings

We use Qwen2.5-VL-7B [35] as the backbone with the visual branch frozen; only the LLM and detransformer are updated. The SFT dataset contains 210K samples with paired auxiliary images from Monet-SFT [45] and VInteraction [34], with geometric alignment (SIFT [28]) and necessity filtering applied. The RL dataset contains 10K samples from ThymeRL [60] and WeMath [33]. We evaluate on V*Bench [51], HRBench4K/8K [48], MME-RealWorld-Lite [61], and VisualPuzzles [39] (OOD), following the `lmms-eval` pipeline [58]. Baselines include: the Qwen2.5-VL-7B backbone, DeepEyes [63](a representative "think with images" approach), LVR [23], Monet [45](two recent works on latent visual reasoning), GPT-4o [19], GPT-5.2 [38], and Gemini-3.1Pro [10]. Full hyperparameters are in Appendix G.

4.2 Main Results

Overall Performance. As shown in Table 1, SCOLAR achieves state-of-the-art among open-source models across four real-world benchmarks. Two conclusions emerge clearly from the results:

Full-resolution auxiliary tokens are critical for fine-grained spatial understanding. SCOLAR's advantage over prior latent reasoning methods is most pronounced on Fine-grained Cross-instance Perception (FCP) metrics—the sub-task that directly requires preserving spatial details from high-resolution inputs. This confirms our core design: by generating auxiliary tokens at input resolution in

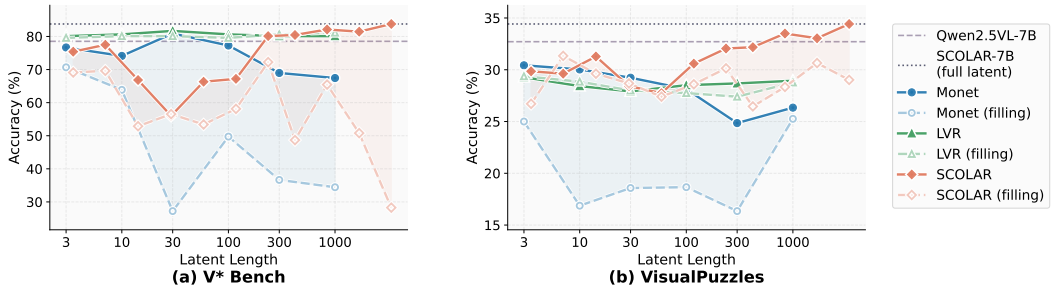


Figure 4: **Latent Length Scaling & Meaningless Padding Experiments on V* Bench and VisualPuzzles.** Solid lines: normal inference; dashed lines: latent tokens replaced with meaningless padding; shaded regions: performance gap.

a single shot, each token independently reconstructs a distinct spatial region, collectively providing the LLM with substantially richer local context than autoregressive methods whose information gain collapses along the chain.

Scalable latent reasoning bridges the gap between 7B open-source and closed-source models. SCOLAR-7B surpasses GPT-4o on V* Bench by over 16 points and GPT-5.2 on MME-RealWorld-Lite by nearly 10 points. This demonstrates that, when latent length can be effectively scaled without degradation, even a 7B model can match or exceed much larger systems—the key is ensuring that additional latent tokens contribute genuine visual semantics rather than redundant context.

Table 2: **Performance on VisualPuzzles (OOD).**

Model	VisualPuzzles					
	Overall	Algo.	Anal.	Dedu.	Indu.	Spat.
GPT-4o [19]	41.3*	49.2*	58.3*	49.0*	27.3*	26.2*
Gemini-3.1Pro [10]	47.95	60.69	61.14	52.00	28.23	38.11
Qwen2.5-VL-7B [1]	32.71	37.02	21.80	47.50	26.32	21.80
Deepeyes [63]	32.96	37.79	27.01	41.00	26.79	27.01
Pangea-7B [57]	31.3*	32.4*	23.7*	38.5*	28.7*	**32.5***
Monet [45]	**35.02***	**45.80***	**30.81***	47.50*	26.79*	25.52*
SCOLAR(SFT only)	32.88	37.79	29.38	44.50	28.71	25.87
SCOLAR-7B	34.42	33.59	29.86	**49.00**	**29.67**	31.82
Δ (vs Qwen2.5VL-7B)	+1.71	-3.43	+8.06	+1.50	+3.35	+10.02

Out-of-Distribution Generalization. As shown in Table 2, SCOLAR-7B generalizes well to VisualPuzzles—a benchmark entirely absent from its training distribution. The improvements are concentrated in sub-tasks that most heavily rely on visual spatial relationships: *spatial reasoning* and *analogical reasoning* show the largest gains over the backbone, while *deductive reasoning* achieves the best result among all open-source models. This pattern is consistent with SCOLAR's design: the detransformer generates auxiliary tokens that encode transferable spatial structure from the original image, rather than task-specific patterns memorized from training data. The fact that these gains emerge on abstract visual puzzles—a domain with no overlap with the SFT data—confirms that the auxiliary reasoning pathway captures generalizable visual structure, not domain-specific shortcuts.

4.3 Latent Length Scaling and Meaningless Padding Experiments

To systematically evaluate the scalability of existing latent reasoning methods and validate the semantic validity of their latent tokens, we conduct two complementary experiments on V* Bench and VisualPuzzles with Latent Length (3 / 10 / 30 / 100 / 300 / 1000 and full latent) as the x-axis. For Monet, we evaluate at different latent size settings. For SCOLAR, we control the length of auxiliary visual token sequences via the **pooling ratio** applied after detransformer output: a smaller pooling ratio yields a longer latent sequence with more complete visual spatial resolution preserved; full latent corresponds to no pooling. In the **padding experiment**, we additionally replace each method's latent tokens with fixed meaningless padding characters at inference time, forcing the model to bypass real latent reasoning.

Length Scaling Patterns (Solid Lines). As Latent Length increases from 3 to 1000, Monet's accuracy monotonically decreases on both benchmarks, confirming that autoregressive information gain collapse directly translates into performance degradation at longer sequences. LVR remains broadly stable but fails to benefit from longer latent chains—consistent with the finding that its tokens carry negligible semantic content regardless of length.

SCOLAR's curve shows a distinctive **three-phase pattern**: at extreme pooling (Latent Length ≤ 10), performance stays high as the model relies mainly on text reasoning; at moderate pooling ($\approx$64$\times$–

256×), accuracy drops sharply because auxiliary tokens are numerous enough to affect attention but too heavily pooled to carry real semantics—solid and dashed lines nearly converge, confirming semantic emptiness; at low-to-no pooling (Latent Length ≥ 100), performance continuously recovers and reaches its peak at full latent. This three-phase pattern reveals a critical insight: *excessively pooled visual features have semantically degraded to meaningless noise*, making them ineffective as alignment targets regardless of generation paradigm—the pooling ratio, not merely the generation method, determines whether auxiliary tokens carry useful information.

Meaningless Padding Analysis (Dashed Lines). The padding experiment reveals the semantic validity of each method's latent tokens by intervening on their content while preserving their structural presence. For LVR, replacing latent tokens with meaningless padding produces performance nearly identical to normal inference on both benchmarks, with solid and dashed curves almost perfectly overlapping at all lengths. This indicates that LVR learned a shortcut of "reasoning with meaningless context"—its generated tokens carry no effective visual semantics, and the model never extracted useful information from them in the first place. For Monet, padding causes sharp performance drops, indicating its latent tokens have structural impact on decoding. However, combined with its systematic degradation as chain length grows, this suggests Monet's dependence is more on the "structural presence" of latent tokens than on progressively richer visual semantics—this structural coupling introduces negative interference when sequences grow longer. For SCOLAR, replacing auxiliary visual tokens with meaningless padding causes significant and *increasing* degradation as sequences grow longer (wider shaded regions at higher Latent Length). This pattern is precisely what we expect from semantically valid tokens: longer sequences encode more genuine visual information, so their removal causes proportionally greater loss. This independently validates that SCOLAR's self-consistent generation paradigm produces auxiliary tokens with real, usable visual semantics.

4.4 Analysis of Latent Token Information Gain.

To directly explain *why* autoregressive latent generation fails at longer sequences, we quantify the incremental information contributed by each new latent token using **Orthogonal Projection Residuals** [41]. Intuitively, IG_t measures the ℓ_2-norm of the component in the $(t{+}1)$-th token that cannot be linearly reconstructed from the preceding t tokens (formal definition in Appendix D). Figure 5 reports the median IG_t (with Inter-Quartile Range band) on a log scale.

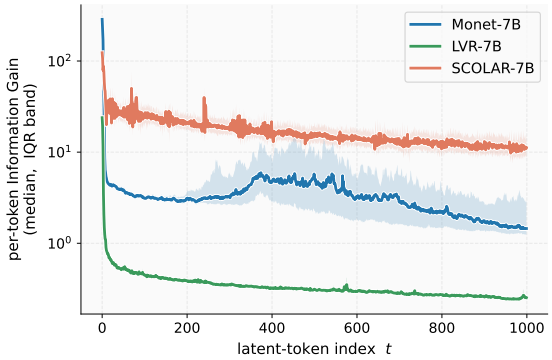


Figure 5: **Step-wise Information Gain of Latent Tokens for Each Method** (Orthogonal Projection Residuals, log y-axis)

Results and Analysis As shown in Figure 5, LVR's IG_t drops to near zero after $t = 10$—subsequent tokens are purely redundant, echoing the padding experiment where LVR is immune to meaningless tokens. Monet's IG_t is higher but continuously decays, confirming the inherent limitation of autoregressive generation: each step depends heavily on prior outputs, so "purely new" features decay exponentially with chain length.

In contrast, SCOLAR maintains IG_t ≈10–30 across all 1000 positions with only gradual decay, because the detransformer maps auxiliary tokens directly from full-sequence hidden states in a single shot—each token is constructed independently of other latent tokens, continuously introducing new visual features that cannot be linearly reconstructed from preceding tokens.

4.5 Ablation Studies

To validate the necessity of each design decision, we systematically remove or replace key components in Table 3. The ablation reveals four design insights (detailed analysis in Appendix E):

(1) Genuine reasoning dependency on auxiliary content. Replacing auxiliary tokens with meaningless placeholders causes *catastrophic* collapse—far worse than simply removing them—confirming

Table 3: **Ablation study.** Each row removes or replaces one component from the full SCOLAR-7B pipeline. Overall accuracy (%) is reported on five benchmarks.

Model	V*	HRBench4K	HRBench8K	MME-RW-Lite	VisualPuzzles
SCOLAR-7B (full)	83.77	75.50	67.63	59.87	34.42
+ disable auxiliary	78.53	71.75	64.12	57.17	31.25
+ replace auxiliary with placeholders	24.61	27.75	31.25	24.60	27.23
+ replace auxiliary with raw visual emb	76.96	72.88	65.00	54.61	28.51
SCOLAR-SFT (w/o ALPO)	84.29	72.50	65.25	53.62	32.88
+ GRPO (replace ALPO)	82.72	73.50	66.88	56.38	33.22
w/o Stage 1 (pretraining)	47.64	43.50	47.00	38.30	17.55
w/o token-level loss (at Stage 2)	82.72	71.25	66.12	48.41	32.11
w/o teacher forcing (at Stage 3)	81.68	70.50	65.88	48.88	32.96
w/o Stage 3	78.53	69.75	65.62	43.77	29.97

the LLM relies on their visual semantics, not merely structural presence. Substituting with raw visual embeddings also degrades results, showing the detransformer's context-aware transformation is essential.

(2) Detransformer pretraining is prerequisite. Without Stage 1, an untrained detransformer injects incoherent signals that corrupt the LLM's visual context, causing catastrophic failure.

(3) Teacher-forcing annealing and token-level weighting address complementary failures. Removing Stage 3 annealing disproportionately hurts benchmarks requiring sustained visual reasoning (MME-RealWorld-Lite), confirming distribution mismatch is most damaging in long-latent settings. Token-level weighting in Stage 2 ensures reliable trigger learning.

(4) ALPO teaches *when* to activate latent reasoning. Replacing ALPO with standard GRPO yields -3.49% on MME-RealWorld-Lite; ALPO's gains concentrate on deep reasoning tasks (HRBench4K FCP +7.25%, MME-RealWorld-Lite +6.25%) while simpler perception remains stable, confirming r_{aux} selectively encourages auxiliary generation where it genuinely helps.

4.6 Summary of Findings

Three sets of experiments converge on a unified picture of why and how SCOLAR succeeds:

Full-resolution auxiliary tokens are the prerequisite for effective latent scaling. Autoregressive latent methods either degrade (Monet) or stagnate (LVR) with increasing latent length, while SCOLAR's three-phase curve demonstrates that genuine visual resolution—not merely longer sequences—drives performance recovery (Figure 4). The padding experiment further confirms that SCOLAR's tokens carry genuine, length-proportional visual semantics.

Single-shot generation prevents information decay across the latent sequence. LVR's IG_t collapses after $t = 10$ and Monet's decays monotonically, whereas SCOLAR sustains $IG_t \approx$10–30 across 1000 tokens (Figure 5)—directly attributable to the non-autoregressive paradigm that prevents inter-token dependency from eroding new information.

The LLM forms genuine reasoning dependency on auxiliary semantic content. Disabling auxiliary tokens causes moderate degradation, but replacing them with meaningless placeholders triggers catastrophic collapse (Table 3)—corroborating the padding experiment and confirming SCOLAR reasons over genuine visual semantics, not mere structural presence.

5 Conclusion

We identify **Information Gain Collapse**—caused by autoregressive dependency and excessively pooled supervision targets—as the root cause of degradation in existing latent visual reasoning methods. We propose **SCOLAR**, whose lightweight detransformer generates auxiliary visual tokens in a single shot with each token independently anchored to the original visual space. Combined with three-stage SFT and ALPO, SCOLAR extends acceptable latent CoT length by over 30$\times$ and achieves state-of-the-art on real-world benchmarks with strong OOD generalization.

Limitations. The detransformer's supervision relies on paired input–auxiliary image data; ALPO's optimization of detransformer parameters remains implicit. Future work will explore unsupervised

settings, explicit end-to-end gradient signals for the continuous visual space, multi-step iterative visual reasoning, and improved interpretability of auxiliary visual tokens.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. Soft tokens, hard truths. *arXiv preprint arXiv:2509.19170*, 2025.

[3] Jiawei Chen, Zhe Chen, Chaoqun Du, Maokui He, Wei He, Hengtao Li, Qizhen Li, Zide Liu, Hao Ma, Xuhao Pan, et al. Streamingclaw technical report. *arXiv preprint arXiv:2603.22120*, 2026.

[4] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[5] Wei Chen, Chaoqun Du, Feng Gu, Wei He, Qizhen Li, Zide Liu, Xuhao Pan, Chang Ren, Xudong Rao, Chenfeng Wang, et al. Mindgpt-4ov: An enhanced mllm via a multi-stage post-training paradigm. *arXiv preprint arXiv:2512.02895*, 2025.

[6] Yang Chen, Yufan Shen, Wenxuan Huang, Shen Zhou, Qunshu Lin, Xinyu Cai, Zhi Yu, Botian Shi, and Yu Qiao. Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback. *arXiv preprint arXiv:2507.20766*, 2025.

[7] Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025.

[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.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[10] Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026. Accessed: 2026-05-01.

[11] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. In *ICML*, 2025.

[12] Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

[13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

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

[16] Zefeng He, Xiaoye Qu, Yafu Li, Tong Zhu, Siyuan Huang, and Yu Cheng. Diffthinker: Towards generative multimodal reasoning with diffusion models, 2025.

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

[18] Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection. *arXiv preprint arXiv:2505.20289*, 2025.

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[20] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[21] Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. *arXiv preprint arXiv:2509.12132*, 2025.

[22] Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *arXiv preprint arXiv:2505.16192*, 2025.

[23] 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.

[24] 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. *arXiv preprint arXiv:2408.03326*, 2024.

[25] 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.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[28] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[29] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[30] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[31] 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.

[32] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: A visual language model with chain-of-manipulations reasoning. In *ICLR*, 2025.

[33] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[34] Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, Peiqing Yang, Shiqiang Lang, Enhui Wan, Xiaowan Wang, Yida Xu, Lan Yang, et al. V-thinker: Interactive thinking with images. *arXiv preprint arXiv:2511.04460*, 2025.

[35] Qwen Team. Qwen2.5 technical report, 2025.

[36] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[37] Yang Shi, Jiaheng Liu, Yushuo Guan, Zhenhua Wu, Yuanxing Zhang, Zihao Wang, Weihong Lin, Jingyun Hua, Zekun Wang, Xinlong Chen, et al. Mavors: Multi-granularity video representation for multimodal large language model. *arXiv preprint arXiv:2504.10068*, 2025.

[38] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[39] Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

[40] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

[41] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[43] Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

[44] Jianwei Wang, Ziming Wu, Fuming Lai, Shaobing Lian, and Ziqian Zeng. Synadapt: Learning adaptive reasoning in large language models via synthetic continuous chain-of-thought. *arXiv preprint arXiv:2508.00574*, 2025.

[45] 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. In *CVPR*, 2026.

[46] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025.

[47] 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.

[48] 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.

[49] Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025.

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

[51] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

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

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

[54] 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.

[55] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

[56] En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

[57] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024.

[58] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

[59] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms. *arXiv preprint arXiv:2505.15436*, 2025.

[60] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

[61] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

[62] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025.

[63] 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.

[64] Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025.

A Detransformer Architecture Details

This section provides the full architectural specification of the detransformer module introduced in Section 3.1.

Core Architecture. The detransformer D_ϕ is a two-layer Transformer that operates on the full-sequence hidden states $H \in \mathbb{R}^{L\times d}$ extracted from an intermediate LLM layer ℓ. Its computation proceeds as:

$$Z_0 = \mathrm{LN}(H), \tag{10}$$
$$Z_{m+1} = \mathrm{MSA}(Z_m) + Z_m, \tag{11}$$
$$Z'_{m+1} = \mathrm{FFN}(\mathrm{LN}(Z_{m+1})) + Z_{m+1}, \quad \text{for } m = 0, 1, \ldots, M-1. \tag{12}$$

The output of the final layer directly serves as the auxiliary visual editing tokens: $A_0 = Z'_M$. In practice, $N_a = N_v$, i.e., the detransformer produces one auxiliary token for each input visual token, preserving spatial structure.

Gated Residual Fusion. The auxiliary tokens are fused with original visual features via gated residual addition:

$$A = V + g(A_0), \qquad g(A_0) = \alpha\, A_0, \quad \alpha \in [0, 1], \tag{13}$$

where V denotes the original visual features and α controls the editing strength. This residual design keeps auxiliary features anchored to the original visual context and prevents degenerate edits.

Design Rationale. Self-attention over the *full* sequence H (not just visual positions) is critical: it allows the detransformer to modulate auxiliary visual generation based on textual reasoning context—the question semantics, Phase-1 reasoning, and trigger token all participate in attention, guiding *what* visual information to reconstruct. In practice, $M = 2$ layers provide sufficient capacity (see preliminary study in Appendix H).

B SFT Training Data Construction

Training requires image–auxiliary-image–answer interleaved triples. We apply a three-step data pipeline to ensure quality:

Step 1: Geometric Alignment. SIFT [28, 29] feature matching aligns and crops the auxiliary image to the input image, enforcing spatial consistency. Pairs where feature matching fails are discarded.

Step 2: CoT Distillation. We use GPT-4o to generate structured Phase-1 reasoning chains $y^{(1)}$ given $(I, x, I_{\text{aux}}, y^{(2)})$, covering: preliminary visual analysis, judgment of whether finer visual information is needed, and concluding with the `<auxiliary>` trigger. Richer Phase-1 context directly benefits the detransformer, which conditions on the LLM hidden states encoding this text.

Step 3: Necessity Filtering. We retain only samples where accuracy significantly improves when auxiliary information is provided versus withheld, ensuring the model learns to trigger auxiliary generation only when genuinely beneficial.

The final dataset contains 210K samples from Monet-SFT [45] and VInteraction [34].

C ALPO: Design Rationale and Implicit Synergy

Why Standard GRPO Is Insufficient. As discussed in Section 3.3, SCOLAR inserts continuous auxiliary embeddings $e_{\text{aux}} \in \mathbb{R}^{N_{\text{vis}}\times d}$ between Phase 1 and Phase 2 generation. These have no discrete distribution and cannot participate in standard importance sampling ratios $r_t(\theta) = \pi_\theta(o_t \mid \cdot)/\pi_{\theta_{\text{old}}}(o_t \mid \cdot)$. ALPO resolves this by caching e_{aux} during rollout and treating it as fixed context during the policy update, restricting gradient computation to discrete tokens only.

Implicit Cooperative Optimization. Although the detransformer D_ϕ does not directly receive RL gradients, a positive feedback loop emerges: as ALPO improves the text policy, Phase-1 generation

becomes more informative, producing richer hidden states H as input to D_ϕ; the detransformer, conditioned on improved H, generates more targeted auxiliary features; these in turn enable better Phase-2 answers, reinforcing the cycle. The ablation in Table 3 provides indirect evidence: ALPO improves MME-RealWorld-Lite by +6.25% over SFT-only (53.62%→59.87%), with gains concentrated on deep reasoning sub-tasks (HRBench4K FCP: 61.25%→68.50%), confirming that r_{aux} effectively guides the model to activate and benefit from the auxiliary path where it matters most.

D Information Gain Measurement

To quantify the incremental information contributed by each new latent token, we use orthogonal projection residuals [41]. Let the embedding matrix of the first t latent tokens be $\mathbf{E} \in \mathbb{R}^{t\times d}$, and the $(t{+}1)$-th token embedding be $\mathbf{e}_{\text{new}} \in \mathbb{R}^d$. The information gain is:

$$IG_t = \left\| \mathbf{e}_{\text{new}} - \mathbf{E}^\top (\mathbf{E}\mathbf{E}^\top)^{-1} \mathbf{E}\, \mathbf{e}_{\text{new}} \right\|_2, \tag{14}$$

which measures the ℓ_2-norm of the component in $\mathbf{e}_{\text{new}}$ that cannot be linearly reconstructed from the preceding t tokens. For sequences with $t > 500$, we use SVD low-rank approximation (retaining top-k singular vectors, $k = \min(t, 256)$) to reduce cost from $O(t^2 d)$ to $O(k \cdot t \cdot d)$.

Figure 5 reports the median IG_t with inter-quartile range across the test set on a log scale.

E Detailed Ablation Analysis

Table 3 reports ablations from two baselines: (1) full SCOLAR-7B (with ALPO) for evaluating auxiliary token utility, and (2) SCOLAR-SFT (all three stages, without ALPO) for evaluating training stage contributions. We summarize the key findings below.

Table 4: Ablation insights summary. Δ denotes drop from the respective baseline.

Ablation	**Largest Δ**	**Insight**
Disable auxiliary	-5.24 (V*)	Auxiliary tokens contribute genuine reasoning value, not merely extended context.
Replace with placeholders	-59.16 (V*)	Catastrophic collapse confirms the LLM depends on *semantic content*, not structural presence of auxiliary tokens.
Replace with raw visual emb	-6.81 (V*)	Context-aware transformation by the detransformer is essential; simply repeating input features is insufficient.
w/o Stage 1	-36.65 (V*)	An untrained detransformer injects incoherent signals, causing catastrophic failure. Pretraining is prerequisite.
w/o token-level loss	-5.21 (MME)	Weighted NTP ensures reliable trigger learning; without it, the model under-triggers on hard samples.
w/o teacher forcing	-4.74 (MME)	Distribution mismatch is most damaging on benchmarks requiring sustained visual reasoning.
w/o Stage 3	-9.85 (MME)	Joint reasoning is the critical stage for learning to exploit self-generated auxiliary tokens.
GRPO (replace ALPO)	-3.49 (MME)	ALPO's r_{aux} selectively encourages auxiliary activation where it helps; GRPO lacks this mechanism.

Semantic validity (rows 1–3). The contrast between "disable" (-5.24 on V*) and "replace with placeholders" (-59.16) is striking: removing auxiliary tokens causes moderate degradation, but injecting *semantically empty* tokens in their place triggers catastrophic confusion. This confirms the LLM has learned to deeply condition its reasoning on auxiliary content—when that content is meaningless noise, it actively misleads rather than merely being ignored.

Training stages (rows 4–7). The ordering of severity—Stage 1 $\gg$ Stage 3 $>$ teacher-forcing $\approx$ token-weighting—reveals a clear dependency chain: basic visual alignment (Stage 1) is the foundation; without it, no subsequent stage can function. Stage 3's joint training has the next-largest impact because it is where the model learns to reason over its own generated features rather than ground-truth substitutes.

Table 5: **Complete per-category results on MME-RealWorld-Lite.**

Model	Overall	Reasoning					Perception					
		Overall.	Monitor.	Auto. Driv.	OCR	Diag.&Tab.	Overall.	Monitor.	Auto. Driv.	OCR	Diag.&Tab.	Remote Sens.
GPT-5.2 [38]	50.23	45.87	40.67	38.50	68.00	61.00	53.04	37.62	38.86	79.60	83.00	54.67
Qwen2.5-VL-7B [35]	42.73	35.87	28.67	23.25	72.00	61.00	47.13	28.53	28.57	86.00	83.00	41.33
DeepEyes [63]	54.25*	50.53*	46.67*	40.25*	78.00*	70.00*	56.63*	43.89*	38.86*	90.00*	84.00*	51.33*
Monet [45]	55.50*	51.07*	46.00*	41.50*	73.00*	**75.00***	58.34*	41.07*	48.86*	85.60*	84.00*	54.67*
SCOLAR (SFT only)	53.62	48.00	50.67	40.00	64.00	60.00	57.23	47.34	49.43	80.00	72.00	48.67
SCOLAR-7B	**59.87**	**54.80**	**61.33**	**46.25**	**77.00**	57.00	**63.13**	**50.47**	**55.14**	**90.00**	**75.00**	**56.00**

ALPO vs GRPO (row 8). Both improve over SFT-only, but ALPO's gains concentrate on tasks requiring deep visual reasoning (HRBench4K FCP: +7.25%, MME-RealWorld-Lite: +6.25%) while V*Bench sees a slight trade-off (-0.52). This pattern confirms that r_{aux} specifically incentivizes auxiliary activation on hard samples rather than uniformly shifting the policy.

F Full Results on MME-RealWorld-Lite

Table 5 presents per-category results. SCOLAR achieves consistently strong performance across all categories including monitoring, autonomous driving, OCR, and diagram understanding.

G Experimental Setup

Training Details. The visual branch (encoder + projector) is frozen; only the LLM and detransformer are updated. The detransformer takes as input the hidden states at the input of layer $\ell = 20$ (i.e., the output of layer 19) of the 28-layer LLM. Full hyperparameters are listed in Tables 6a and 6b.

Table 6: **Training hyperparameters.**

(a) SFT

Hyperparameter	Value
Stage-1 learning rate	1e-4
Stage-2 learning rate	1e-5
Stage-3 learning rate	4e-6
Batch size	1
Gradient accumulation steps	16
World size (GPUs)	8
Optimizer	AdamW [27]
Max sequence length	16384
Max pixels	4194304
Max visual tokens	6000
Selected hidden layer ℓ	20
Visual loss type	MSE + cosine similarity
Cosine loss weight	0.5
Visual loss weight λ	2
Teacher-forcing annealing	$1 \rightarrow 0$ over 700 steps
SFT Stage 1 steps	3000 ($\approx$2 epochs)
SFT Stage 2 steps	100
SFT Stage 3 steps	1500 ($\approx$1 epoch)

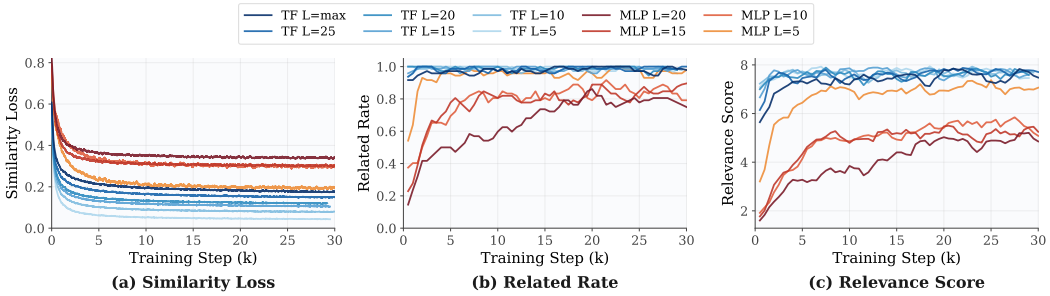
(b) RL (ALPO)

Hyperparameter	Value
Learning rate	4e-6
Batch size	256
Weight decay	0.01
Rollout group size G	8
Temperature	1.0
Max response length	4096
KL coefficient	1e-2
Max pixels	1003520
Optimizer	AdamW
Auxiliary reward weight λ_{aux}	0.3
Training steps	54

Compute Resources. SFT Stage 1: $\approx$12 H200 GPU-hours per epoch. Stages 2 and 3: $\approx$50 H200 GPU-hours per epoch each. RL (ALPO): $\approx$800 H200 GPU-hours per epoch.

Evaluation. All benchmarks are evaluated with `lmms-eval` [58] using default settings. V*Bench [51] and HRBench [48] measure fine-grained visual perception; MME-RealWorld-Lite [61] covers real-world reasoning and perception; VisualPuzzles [39] tests abstract visual logic (OOD).

Baselines. We compare against: (1) Qwen2.5-VL-7B backbone [35]; (2) DeepEyes [63] ("think with images" via cropping); (3) LVR [23] and Monet [45] (latent visual reasoning); (4) Proprietary models: Pangea-7B [57], GPT-4o [19], GPT-5.2 [38], and Gemini-3.1Pro [10].



(a) Similarity Loss **(b) Related Rate** **(c) Relevance Score**

Figure 6: **Preliminary study on detransformer architecture and hidden layer selection.** Transformer (TF, blue) vs. MLP (red/orange) at different extraction layers. **(a)** Similarity loss. **(b)** Related Rate. **(c)** Relevance Score (1–10).

H Hidden Layer Selection

Selection of layer ℓ. The detransformer operates on hidden states from an intermediate layer ($\ell = 20$, i.e., the input to the 20th layer out of 28). This balances three considerations: (1) final-layer hidden states are optimized for next-token prediction and risk collapsing the LLM's text semantics; (2) final layers retain less visual information, making reconstruction harder; (3) too shallow a layer lacks sufficient computational depth for meaningful visual editing.

In a preliminary study (Figure 6), we compared a two-layer Transformer detransformer against a two-layer MLP for visual feature reconstruction from hidden states at layers $\ell \in \{5, 10, 15, 20, 25, \max\}$. We evaluate on 60 samples from LLaVA-in-the-Wild [26], using Qwen2.5-VL-235BA22B as a judge to assess semantic relatedness between reconstructed tokens and original images. Two metrics are reported: *Related Rate* (proportion judged as related) and *Relevance Score* (average rating, 1–10).

The Transformer architecture significantly outperforms MLP across all layers. At intermediate-to-deep layers ($\ell \geq 10$), the Transformer achieves Related Rate $\approx$1.0 and Relevance Score 7–8, confirming sufficient capacity for decoding visual semantics from hidden states.

I Additional Limitations and Future Works

Beyond those discussed in the main text, we identify three limitations of the current system and outline corresponding directions for future research.

(1) Incomplete layer selection ablation. A comprehensive ablation over all layer choices ℓ within the full training pipeline was not completed due to the prohibitive computational cost (each choice requires retraining the entire three-stage SFT + RL pipeline). We consider this a valuable direction for future work: such a study would shed light on the residual visual semantic information preserved at each layer of large language models, contributing to the broader understanding of LLM interpretability.

(2) KV cache recomputation during auxiliary injection. After the detransformer generates auxiliary visual tokens, they are appended to the existing context for Phase-2 generation. Currently the Phase-1 KV cache is discarded and the full sequence is recomputed, introducing additional latency. Since auxiliary tokens are *added* rather than replacing existing context, the Phase-1 cache remains valid in principle. A promising direction is to enable the detransformer to read the Phase-1 KV cache, compute KV entries only for the new auxiliary tokens, and append them to the existing cache—eliminating redundant recomputation in Phase-2.

(3) Implicit optimization of the detransformer in ALPO. Currently, ALPO optimizes the detransformer only implicitly—as the text policy improves, it produces richer hidden states that indirectly benefit auxiliary generation, but there is no guarantee of convergence to the optimal auxiliary generation strategy. A more direct approach would be to incorporate the auxiliary token generation into an explicit loss path for D_ϕ, for example by backpropagating reward signals through the continuous auxiliary embeddings to the detransformer parameters. Due to time constraints, this end-to-end optimization was not explored in the current work and is left for future investigation.