ADAPTIVE MULTIMODAL REASONING WITH VISUAL IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models increasingly rely on textual reasoning to solve advanced multimodal tasks. Despite its effectiveness, this paradigm faces two fundamental limitations. First, relying on textual reasoning alone can limit perceptual ability in vision-centric questions, where accurate answers depend on interpreting task-relevant visual evidence. Second, applying textual reasoning uniformly across inputs overlooks question-dependent variation in both reasoning modality and depth, often producing verbose and unnecessary traces when little logical inference is required. To this end, we propose **A**daptive **V**isual-**T**extual **R**easoning (AVTR), which introduces visual imagination as a complementary reasoning modality to textual thinking, enabling the model to revisit its visual observations and distill the salient evidence through generative visual reflection. Furthermore, we formulate reasoning-modality selection as a learned routing problem, allowing a single policy to adaptively harness the complementary strengths of visual reflection and textual reasoning in response to the perceptual and inferential demands of the input. Across 16 benchmarks, AVTR improves the average score by 4.5 points over Qwen3.5 baseline, while reducing inference latency by 33%. It also scales effectively into a smaller size backbone, outperforming the evaluated 2B-scale baselines. Our code will be made publicly available.

1 INTRODUCTION

Vision-Language Models (VLMs) (Liu et al., 2023; Zhu et al., 2024; Wang et al., 2025a; Bai et al., 2025) are multimodal in what they observe but largely unimodal in how they reason. Although they perceive both visual and textual inputs, their reasoning process is predominantly carried out in autoregressive text tokens (Wei et al., 2022; Zhang et al., 2023; Shao et al., 2024a). This paradigm is effective when relevant visual evidence has been captured, leaving only logical inference to be performed. However, in practice, the perceived visual input can be either incomplete or unfaithful, especially on challenging vision tasks. As an example, Qwen3.5-9B (Qwen Team, 2026c) achieves only 34.7% accuracy on MMSI-Bench (Yang et al., 2026a), substantially below the 97% human accuracy, while textual reasoning raises the score by only 0.3 points to 35.0%. Such marginal gain from textual reasoning suggests that visual perception remains a key bottleneck, constraining what further reasoning can achieve with incomplete or unreliable visual evidence.

Several approaches have been explored in prior work to address the perceptual bottleneck. Explicit visual-tool methods zoom, crop, or augment the input, but require external execution and additional visual processing at inference (Gupta & Kembhavi, 2023; Surís et al., 2023; Hu et al., 2024; Yang et al., 2023; Shen et al., 2025; Zheng et al., 2026; Su et al., 2026). Unified models combine understanding and generation (Dong et al., 2024; Zhou et al., 2025a; Xie et al., 2025b), but generation is rarely used to provide useful feedback for perception or reasoning, a direction explored only recently (Li et al., 2025; Song et al., 2026). Latent visual-reasoning methods extend reasoning beyond discrete text tokens to continuous, semantically aligned visual states (Yang et al., 2026b; Li et al., 2026a; Pham & Ngo, 2025; Wang et al., 2026a; Shen & Elhoseiny, 2026), yet a non-negligible gap remains between latent semantic alignment and faithful, question-conditioned visual perception.

In this work, we present AVTR, which introduces **visual imagination** as an intermediate step to complement textual reasoning. It encourages the model to revisit relevant visual evidence before generating a textual answer, rather than reasoning directly from its initial visual observation. We

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

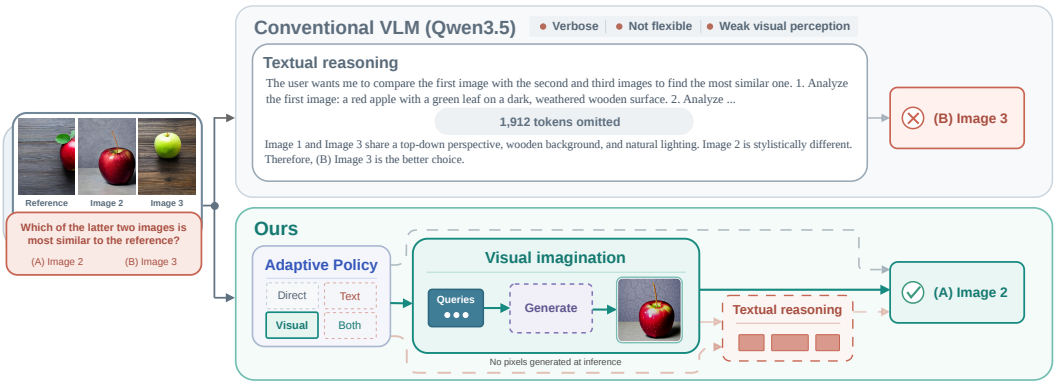

Figure 1: **Adaptive multi-modal reasoning with visual imagination.** Conventional VLMs perceive multimodal inputs but reason entirely in text. AVTR adaptively harnesses the complementary strengths of visual imagination and textual reasoning according to the perceptual and inferential demands of the input. Visual imagination forms a question-conditioned visual state through visual queries with training-time generative supervision, while requiring no pixel generation at inference.

enable this process by bridging the output from a language model to the visual generative space, allowing it to produce visual evidence, which is useful to the task. We further teach the model to distill salient visual evidence through **generative visual reflection** and leverage it for question-answering tasks. More importantly, the visual generator is solely used as the training-time supervision and is completely removed at inference. Therefore, the model is capable of reasoning from the learned visual states without decoding them into pixels.

Furthermore, our empirical analysis reveals that visual imagination and textual reasoning exhibit **complementary strengths** across different reasoning demands (Figure 3). Visual imagination is particularly effective for improving perception-intensive tasks by distilling task-relevant visual evidence, whereas textual reasoning remains a great advantage for explicit relational and compositional inference. This complementarity calls for an adaptive selection of the reasoning route, in a way to best preserve answer correctness while avoiding unnecessary computation. To this end, we propose **adaptive multimodal reasoning**, formulating it as a learned routing problem over visual imagination and textual reasoning. A shared policy selects direct answering, visual imagination, textual reasoning, or their composition according to the demands of each question.

We perform extensive evaluations on various tasks, i.e. high-resolution image understanding, spatial reasoning, video understanding, and temporal grounding over total 16 benchmarks. Across all the benchmarks, AVTR achieves strong average gain over different backbone modes, improving by a large margin of 4.9 points over direct answering and by 4.5 points over textual reasoning. In the meantime, we are able to reduce the average inference latency by 33% w.r.t textual reasoning (Figure 5). Note that particularly on a 2B-parameter backbone, AVTR achieves a significant 6.3 points gain over the reference model. As a result, AVTR demonstrates that both visual imagination and adaptive routing improve reasoning quality and efficiency across different model scales.

Contributions. Our contributions are summarized in threefold:

- We introduce visual imagination as an intermediate reasoning modality that revisits visual observations and distills task-relevant evidence for sophisticated vision tasks.
- We formulate adaptive multi-modal reasoning as a learned routing problem over visual imagination and textual reasoning. Such design enables a single policy to select or compose the two based on each input's perceptual and inferential demands.
- We demonstrate the effectiveness of AVTR on 16 benchmarks spanning diverse reasoning tasks. Our results show consistent improvements in both evaluation metrics and inference efficiency. Meanwhile, it also scales effectively to a smaller 2B backbone.

2 Related Work

Think with Vision. VLMs combine visual encoders with language models and have achieved strong performance on image and video understanding (Liu et al., 2023; Bai et al., 2025; Li et al., 2024a; Tong et al., 2024). Fine-grained visual questions remain challenging because attention can be drawn to visually salient but task-irrelevant tokens (Kang et al., 2025). Existing methods address this by explicitly selecting visual evidence, such as relevant regions or temporal spans, by using tools to crop, zoom, annotate, or re-encode selected regions (Gupta & Kembhavi, 2023; Surís et al., 2023; Hu et al., 2024; Yang et al., 2023; Shen et al., 2025; Zheng et al., 2026; Su et al., 2026). While effective, these approaches typically introduce additional visual processing at inference. AVTR instead forms a question-conditioned visual state that revisits the original observation and distills task-relevant evidence.

Latent Visual Reasoning. Latent reasoning replaces explicit textual traces with continuous states (Hao et al., 2024; Goyal et al., 2024). Recent VLM methods align language-model hidden states with visual features and feed them back as inputs for subsequent reasoning (Yang et al., 2026b; Li et al., 2026a; Pham & Ngo, 2025; Wang et al., 2026a; Shen & Elhoseiny, 2026; Jeon et al., 2026). This output-to-input reuse may introduce a representation gap, while autoregressive state generation adds sequential inference cost. In contrast, AVTR inserts learnable visual queries directly into the VLM input and forms a question-conditioned visual state in a single forward pass through generative visual reflection in latent pixel.

Adaptive Reasoning. Recent work has explored adapting textual reasoning to input difficulty. AdaptThink (Zhang et al., 2025) learns to switch between Thinking and NoThinking modes, while VideoAuto-R1 (Liu et al., 2026) uses the confidence of an initial answer to decide whether further reasoning is needed. Other methods adapt textual computation through stopping or inference-time budgets (Chen et al., 2025b; Sui et al., 2025; Fang et al., 2025; Muennighoff et al., 2025). These approaches primarily control the amount of textual reasoning. AVTR extends adaptive reasoning to multimodal routing, allowing the model to route between visual imagination and textual reasoning according to the perceptual and inferential demands of each question.

3 Method

Given a visual input $\mathcal{V}$ and a question q, a standard VLM π_θ predicts an answer a from the visual and textual context. AVTR augments this process with visual imagination, forming a question-conditioned visual state that distills task-relevant visual evidence for answer generation. **Visual replay** first aligns this state with the visual generative space, **visual reflection** makes it question-dependent, and **imagination-guided** supervised finetuning teaches the VLM to leverage it for answering. We then learn an outcome-guided **adaptive routing policy** from the empirical performance of different reasoning routes. The policy adaptively selects among direct answering, visual imagination, textual reasoning, and their composition for each input. Figure 2 gives an overview.

3.1 Visual Imagination Through Replay and Reflection

Visual Replay. As illustrated in Figure 2(a), visual replay first learns a set of visual states that connect the VLM representation space to a visual generator through reconstruction (Pan et al., 2025). Let $\boldsymbol{Q} \in \mathbb{R}^{n\times d}$ denote n learnable visual queries with hidden dimension d. The text prompt p is either the associated image caption or a general reconstruction instruction, "Please reconstruct this image". The VLM processes $[\mathcal{V}; p; \boldsymbol{Q}]$ in a single forward pass, and $\boldsymbol{H}_Q \in \mathbb{R}^{n\times d}$ denotes the last hidden states at the query positions. A connector g_ψ, parameterized by ψ, maps $\boldsymbol{H}_Q$ to the conditioning space of the visual generator v_ϕ, parameterized by ϕ. We optimize

$$\mathcal{L}_{\text{imag}} = \mathbb{E}\left[\left\|v_\phi(\boldsymbol{z}_\sigma, \sigma, g_\psi(\boldsymbol{H}_Q)) - (\boldsymbol{\epsilon} - \boldsymbol{z}_0)\right\|_2^2\right], \qquad \boldsymbol{z}_\sigma = (1-\sigma)\boldsymbol{z}_0 + \sigma\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{z}_0$ is the latent encoding of the visual target produced by the frozen autoencoder, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ is Gaussian noise, $\sigma \sim \mathcal{U}(0, 1)$ is the flow interpolation step, and $\boldsymbol{z}_\sigma$ is the interpolated latent at step σ. During visual replay, the visual target is the input $\mathcal{V}$ itself.

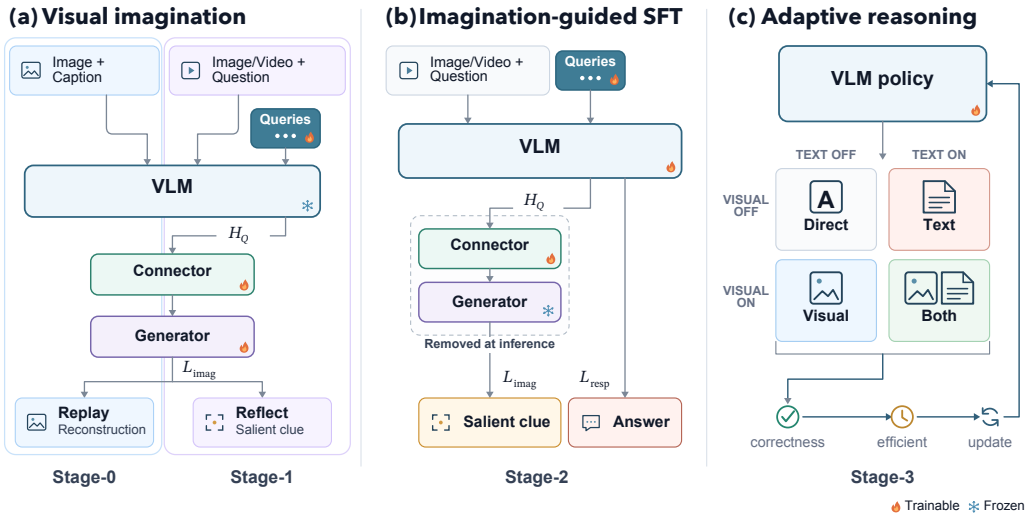

Figure 2: **Method overview. (a) Visual imagination** is learned in two stages with shared trainable parameters but different inputs and supervision: *visual replay* reconstructs the input from visual queries, while *visual reflection* conditions them on the question to generate relevant visual evidence. **(b) Imagination-guided SFT** jointly optimizes visual-generation and response objectives, enabling the resulting visual state to guide answer generation. The connector and generator provide training-only supervision and are removed at inference. **(c) Adaptive reasoning** learns a correctness- and cost-aware policy that selects direct answering, visual imagination, textual reasoning, or visual imagination followed by textual reasoning for each input.

We optimize the visual queries, connector, and generator while keeping the VLM and autoencoder frozen. Since the queries are forward all at once, this stage learns a visual state without autoregressive generation of latent tokens. Reconstructing the visual input from $g_\psi(\boldsymbol{H}_Q)$ encourages $\boldsymbol{H}_Q$ to recover what the model observes as visual input.

Visual Reflection. Visual reflection then reuses the same architecture but replaces generic reconstruction with question-conditioned supervision, as shown in Figure 2(a). We use the question as the prompt, $p = q$, and replace the reconstruction target with a question-relevant visual target $\mathcal{V}_{\text{tgt}}$.

For images, $\mathcal{V}_{\text{tgt}}$ is the salient region or visual evidence of the input visual context; for videos, it is the frame containing the faithful evidence. Training with these targets encourages $\boldsymbol{H}_Q$ to focus on the visual evidence needed to answer the question. Appendix C describes how the targets are localized, verified, and refined.

3.2 Imagination-Guided SFT

Generative alignment makes the query state visually decodable and focuses it on question-relevant evidence, but does not directly supervise its role in answer generation. To bridge this gap, imagination-guided SFT jointly optimizes response supervision and the visual generation objective from the trained generator, as illustrated in Figure 2(b). For samples with a visual target, we use $y = [\texttt{<imagine>}; \boldsymbol{Q}; \texttt{</imagine>}; a]$, where the continuous queries $\boldsymbol{Q}$ are placed before the answer and provide visual evidence for answer generation. These two special tokens are added to the VLM's vocabulary. For those without a target, we use $y = [\texttt{</imagine>}; a]$ and do not require $\boldsymbol{Q}$ or $\mathcal{L}_{\text{imag}}$. Let $\mathcal{T}$ denote the supervised discrete token positions in y, excluding $\boldsymbol{Q}$. We optimize

$$\mathcal{L}_{\text{resp}} = -\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \log \pi_\theta(y_t \mid \mathcal{V}, q, y_{<t}), \qquad \mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{resp}} + \lambda_{\text{imag}} \mathcal{L}_{\text{imag}}, \tag{2}$$

where λ_{imag} controls the weight of the visual generation objective.

The generative loss anchors the query state to the visual target, while the response loss trains the model to leverage this state for answer generation. The aligned generator, connector, and autoencoder are used only for training-time visual supervision and are removed at inference. The visual query states remain in the VLM context and directly condition answer generation.

3.3 Adaptive Multimodal Reasoning

Visual imagination provides visual evidence, whereas textual reasoning supports explicit compositional inference. We let the model decide when to invoke each through two token gates. The visual gate predicts $\ell_v \in \{\texttt{<imagine>}, \texttt{</imagine>}\}$, and the textual gate predicts $\ell_t \in \{\texttt{<think>}, \texttt{</think>}\}$. An opening token activates the corresponding branch, while the closing token skips it. Together, the two gates define the route space $\mathcal{M} = \{\textsc{Direct}, \textsc{Visual}, \textsc{Textual}, \textsc{Visual+Textual}\}$. For VISUAL+TEXTUAL, visual imagination precedes textual reasoning so that the textual trace can attend to the visual state.

Outcome-Guided Gate Learning. For each input $x = (\mathcal{V}, q)$, we force all four routes in $\mathcal{M}$ and sample g continuations under each route, yielding $G = 4g$ rollouts in total. Let $i \in \{1, \ldots, G\}$ index all rollouts, and let $m_i \in \mathcal{M}$ denote the route under which rollout i is sampled. We use $r_i^{\text{ans}} \in \{0, 1\}$ to denote its answer reward. The mean reward of route m is

$$R_m = \frac{1}{g} \sum_{i:\, m_i = m} r_i^{\text{ans}}. \tag{3}$$

Rather than retaining only the best-performing route, we preserve the relative outcomes of all four routes through a soft routing target:

$$p(m \mid x) = \text{Softmax}_{m \in \mathcal{M}} \left(R_m - C(m) \right), \tag{4}$$

where $C(m)$ denotes the route-level cost used in the routing target. We use $C(\textsc{Direct}) < C(\textsc{Visual}) < C(\textsc{Textual}) < C(\textsc{Visual+Textual})$, with $C(\textsc{Direct}) = 0$.

Since the gate tokens are forced rather than sampled during route-complete evaluation, we supervise the routing decisions directly:

$$\mathcal{L}_{\text{gate}} = - \sum_{m \in \mathcal{M}} p(m \mid x) \left[\log \pi_\theta(\ell_v^m \mid x) + \log \pi_\theta(\ell_t^m \mid x, \ell_v^m) \right], \tag{5}$$

where (ℓ_v^m, ℓ_t^m) denotes the pair of gate labels corresponding to route m. The textual gate is evaluated after executing the visual branch specified by ℓ_v^m.

Route-Conditioned Content Optimization. We use the same G rollouts to optimize the remaining reasoning and answer tokens with GRPO (Shao et al., 2024b). We keep the standard GRPO objective and modify only how the group-relative advantage is computed. Rather than centering rewards across all routes, we center each rollout against those sampled under the same route:

$$A_i = r_i^{\text{ans}} - \text{mean}\left(\left\{ r_j^{\text{ans}} : \, m_j = m_i \right\} \right). \tag{6}$$

This prevents differences in route utility from entering the content update, leaving route selection to $\mathcal{L}_{\text{gate}}$. We omit standard-deviation normalization to avoid over-weighting low-variance groups during policy optimization (Liu et al., 2025). Tokens at the two gating positions, together with the continuous visual queries, are masked from the GRPO loss. We denote the resulting loss by $\mathcal{L}_{\text{GRPO}}$. Rollouts with $A_i = 0$ are excluded from $\mathcal{L}_{\text{GRPO}}$, as they provide no relative learning signal within the route. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}}, \tag{7}$$

where λ_{gate} controls the gate objective. The gate learns which route to select, while GRPO learns how to reason within that route.

4 Experiments

4.1 Experimental Setup

Evaluation Benchmarks and Metrics. We evaluate high-resolution image understanding on V* (Wu & Xie, 2024), HR-Bench-4K/8K (Wang et al., 2025b), and BLINK (Fu et al., 2024); spatial reasoning on ERQA (Gemini Robotics Team et al., 2025), MMSI-Bench (Yang et al., 2026a),

Table 1: **High-resolution image understanding and spatial reasoning.**

	HighRes Image				Visual Spatial			
Model	**V***	**HR-Bench-4K**	**HR-Bench-8K**	**BLINK**	**ERQA**	**MMSI-Bench**	**VSI-Bench**	**RealWorldQA**
Proprietary								
GPT-5	72.8	-	-	71.0	65.7	41.9	55.0	76.1
Gemini 2.5 Pro	83.8	87.3	85.4	73.5	55.7	38.0	53.6	68.7
GPT-5 mini	78.6	78.6	74.4	56.7	54.0	34.1	48.7	-
Open foundation models								
Qwen3.5-9B (w/o Think)	86.3	84.7	82.5	70.0	46.0	34.7	57.2	75.4
Qwen3.5-9B (w/ Think)	88.4	88.7	85.1	72.4	52.3	35.0	54.8	80.5
InternVL3.5-8B	70.2	69.7	70.1	59.5	41.0	28.7	56.3	67.5
Cambrian-S-7B	70.6	64.0	53.7	-	39.0	27.1	67.5	64.8
LLaVA-OneVision-2-8B	72.0	72.0	63.1	63.5	43.3	29.6	70.9	69.7
JoyAI-Image-Und-8B	85.9	78.8	76.5	69.6	45.0	35.8	60.1	74.0
AVTR-9B (Ours)	**89.5**	**89.4**	**85.1**	**75.2**	**54.2**	**39.8**	**71.0**	**81.5**

Table 2: **Video understanding and temporal grounding.**

	Video Understanding (Acc)				**Temporal Grounding** (mIoU)			
Model	**MVBench**	**Video-MME**	**LongVideoBench**	**MLVU**	**Charades-STA**	**QVHighlights**	**ActivityNet-Cap.**	**NExT-GQA**
Qwen3.5-9B (w/o Think)	69.1	72.4	67.8	79.1	61.8	67.2	49.8	38.4
Qwen3.5-9B (w/ Think)	71.8	75.4	69.1	80.0	55.2	63.5	42.1	34.4
InternVL3.5-8B	72.1	66.0	62.1	70.2	27.8	31.3	31.3	25.1
Video-R1	65.5	61.8	-	-	-	-	-	-
Time-R1	65.9	54.2	55.8	-	58.8	-	52.1	-
VITAL	-	64.1	-	-	59.9	-	49.8	43.0
VideoChat-R1.5	70.6	65.2	61.4	70.1	60.6	-	35.3	-
VideoAuto-R1-8B	72.0	71.7	67.4	78.1	63.7	61.6	51.9	44.2
LLaVA-OneVision-2-8B	66.2	71.9	68.0	76.6	53.5	63.3	50.5	35.6
AVTR-9B (Ours)	**72.4**	**76.9**	**70.5**	**81.7**	**64.7**	**71.5**	**53.1**	**44.8**

VSI-Bench (Yang et al., 2025a), and RealWorldQA (xAI, 2024); and video understanding on MVBench (Li et al., 2024b), Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025b). These benchmarks are evaluated using their official accuracy metrics. We further evaluate temporal grounding using mIoU on Charades-STA (Gao et al., 2017), QVHighlights (Lei et al., 2021), ActivityNet Captions (Krishna et al., 2017), and NExT-GQA (Xiao et al., 2024). Appendix B provides the baseline sources and complete evaluation protocol.

Implementation Details. We use Qwen3.5-9B and Qwen3.5-2B (Qwen Team, 2026c;b) as the VLM backbones, with $n = 64$ learnable visual queries and a 6-layer connector to SANA-1.6B (Xie et al., 2025a) at 512×512 resolution. Training proceeds in four stages: visual replay on 5M captioned images with the VLM frozen for alignment; visual reflection on 178K question-conditioned targets; SFT on 380K instruction examples; and adaptive routing on 20K questions with $g = 4$ rollouts per route. We set $\lambda_{\text{imag}} = 0.2$ and $\lambda_{\text{gate}} = 0.02$. All models use the same prompts and greedy decoding, with up to 20K visual tokens, 256 video frames sampled at 2 fps, and 8,192 newly generated tokens. Appendix A and B provide the full training and evaluation details.

4.2 Main Results

Image and spatial understanding. Table 1 shows consistent gains across high-resolution understanding and spatial reasoning. Relative to direct answering with the backbone (w/o Think), AVTR improves all eight benchmarks and raises the average from 67.1 to 73.2, a 6.1-point gain. The improvement is particularly pronounced on spatial reasoning (+8.3 points), with the largest gain on VSI-Bench (+13.8 points). This pattern is consistent with the intended role of visual imagination: generating task-relevant visual evidence that supports visual perception and spatial reasoning.

Video understanding and temporal grounding. The same benefit extends from spatial to temporal evidence (Table 2). AVTR improves over direct answering by 3.3 points on average for video understanding and by 4.2 points for temporal grounding, suggesting better localization of task-relevant temporal evidence. We also observe that enabling textual reasoning improves the backbone on video understanding but consistently degrades performance across all four temporal-grounding benchmarks, further motivating adaptive reasoning. For a fair comparison with manageable inference cost, we limit thinking models to a maximum of 8,192 generated tokens rather than adopting the backbone's official 224K long-video setting.

Table 3: **Results with small VLMs.** AVTR uses Qwen3.5-2B as its backbone.

	HighRes Image				Visual Spatial				Video Understanding			
Model	**V***	**HR-Bench-4K**	**HR-Bench-8K**	**BLINK**	**ERQA**	**MMSI-Bench**	**VSI-Bench**	**RealWorldQA**	**MVBench**	**Video-MME**	**LongVideoBench**	**MLVU**
Gemma 4 E2B	55.0	56.5	47.8	45.9	36.3	27.3	21.4	49.7	40.6	56.2	46.4	68.8
MiniCPM-V 4.6 (1.3B)	75.4	71.9	65.6	56.7	39.0	28.0	37.0	63.3	55.8	58.9	55.8	65.2
InternVL3.5-2B	68.1	61.4	58.4	52.1	38.8	25.7	54.0	59.9	**68.2**	57.5	52.1	65.0
Qwen3.5-2B (w/o Think)	82.7	77.1	74.0	48.3	34.0	22.2	48.0	67.7	58.6	60.9	51.9	69.3
Qwen3.5-2B (w/ Think)	81.7	75.4	73.8	60.3	42.8	28.8	31.0	**70.5**	59.6	55.4	49.3	70.2
AVTR-2B (Ours)	**86.4**	**77.6**	**75.8**	**64.4**	**48.0**	**32.7**	**60.8**	69.0	61.3	**62.4**	**59.1**	**72.3**

Effectiveness on a smaller backbone. Table 3 shows that AVTR scales well to the 2B backbone. Across all benchmarks, it improves the average score by 6.3 points over base model, with consistent gains across high-resolution image, spatial, and video understanding. These results demonstrate that adaptive multimodal reasoning remains effective across model scales.

4.3 Visual and Textual Reasoning Are Complementary

Table 4: **Comparison of reasoning modalities.** Under **Think**, ✓ means that a modality is always used, ✗ that it is disabled, and AUTO that the model decides per question.

	Think		HighRes Image				Visual Spatial				Video Understanding		
Model	**Visual**	**Text**	**V***	**HR-Bench-4K**	**HR-Bench-8K**	**BLINK**	**ERQA**	**MMSI-Bench**	**VSI-Bench**	**RealWorldQA**	**MVBench**	**Video-MME**	**LongVideoBench**
Qwen3.5	✗	✗	86.3	84.7	82.5	70.0	46.0	34.7	57.2	75.4	69.1	72.4	67.8
Qwen3.5	✗	✓	88.4	88.7	85.1	72.4	52.3	35.0	54.8	80.5	71.8	75.4	69.1
AVTR	✗	✗	87.9	84.7	82.8	73.3	47.0	37.1	65.8	75.8	70.0	74.3	68.5
AVTR	✓	✗	88.4	86.7	83.8	75.9	48.7	37.4	69.9	76.3	70.3	74.4	69.4
AVTR	✗	✓	89.5	88.3	84.5	73.1	48.2	36.4	63.2	77.1	72.1	74.9	70.1
AVTR	✓	✓	88.5	87.8	83.5	74.2	50.7	37.7	63.5	77.4	70.4	74.4	69.2
AVTR (default)	AUTO	AUTO	89.5	89.4	85.1	75.2	54.2	39.8	71.0	81.5	72.4	76.9	70.5

Reasoning-mode ablation. Table 4 compares different reasoning routes using the same AVTR checkpoint. Enabling visual imagination consistently improves over the direct route. However, no fixed reasoning strategy is uniformly optimal: always using visual imagination, textual reasoning, or their composition is inferior to adaptive routing on average. In particular, simply enabling more reasoning does not guarantee better performance. The benefit therefore comes from selecting the appropriate reasoning modality for each question rather than applying a fixed reasoning pattern globally. Table 7 (Appendix) further analyzes activation rates of the visual and textual gates under the learned adaptive policy for each benchmark.

Complementary strengths of visual and textual reasoning. Figure 3 shows clear modality preferences across representative subtasks. Visual imagination performs better on metric, geometric, and viewpoint-sensitive perception, while textual reasoning is stronger on temporal ordering, counting, and discrete relational reasoning. This complementarity motivates adaptive routing to select the appropriate reasoning modality for each question.

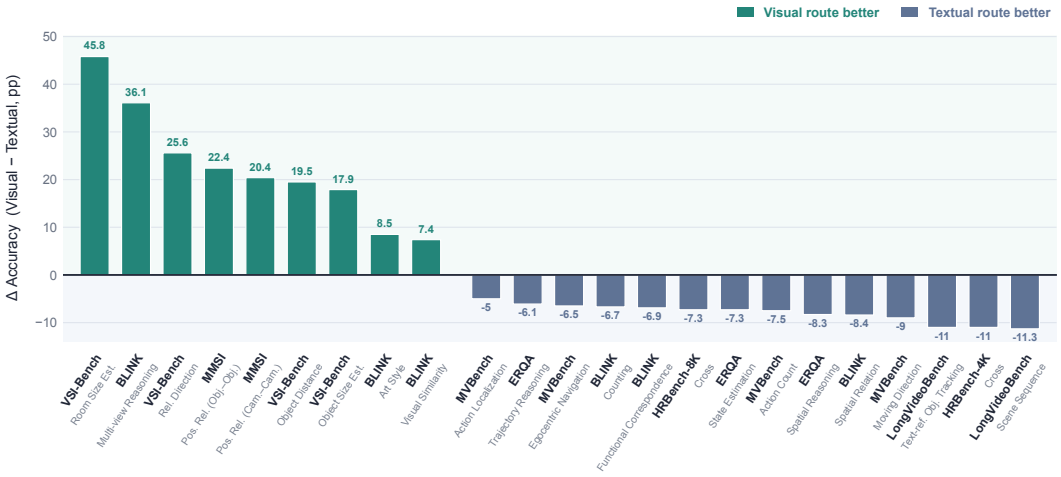

Figure 3: **Complementary strengths of visual and textual reasoning.** Accuracy differences between the visual imagination and textual reasoning on representative subtasks with large performance gaps. Positive values favor visual imagination, while negative values favor textual reasoning.

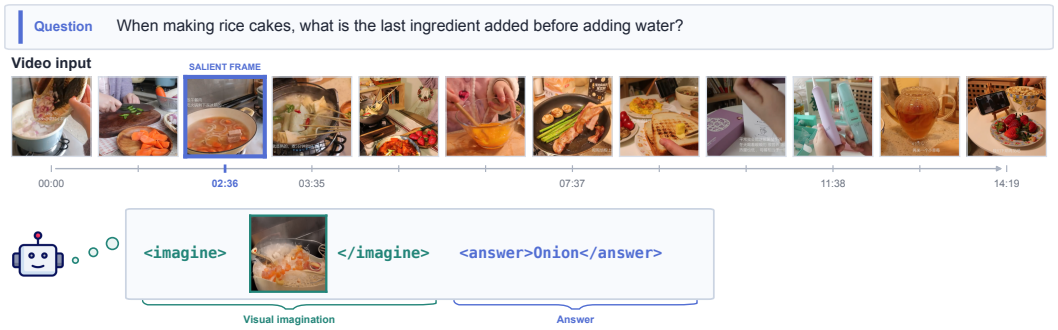

Figure 4: **Visual reflection for long-video understanding.** The rendered image is used only for visualization. The answer is conditioned directly on the continuous visual queries at inference.

Qualitative analysis. Figures 4 and 8 provide qualitative visualizations of the visual reflection results. Figures 9, 10 and 11 further illustrate the complementary strengths of visual and textual reasoning across different examples.

4.4 Ablation Studies

Effect of generative visual supervision. Table 5 first compares our generative supervision with two variants. Pure SFT uses the same instruction data but optimizes only $\mathcal{L}_{\text{resp}}$, while feature-aligned SFT replaces the generative objective with feature-space cosine alignment to the visual target. Both improve over the backbone but remain below AVTR, showing that generative grounding provides a stronger supervision signal for learning task-relevant visual states.

Contribution of each training stage. Table 5 shows that visual replay is critical for establishing the alignment between the VLM and the visual generator. Removing Stage-0 causes the largest degradation, suggesting that later question-conditioned reflection depends on a well-aligned visual

Table 5: **Ablation of visual supervision and training stages.** Pure SFT and feature-aligned SFT control for instruction-tuning data and feature-space supervision, respectively. Each "w/o" variant removes only the named training phase while retaining the remainder of the pipeline.

	HighRes Image				Visual Spatial				Video Understanding		
Variant	**V***	**HR-Bench-4K**	**HR-Bench-8K**	**BLINK**	**ERQA**	**MMSI-Bench**	**VSI-Bench**	**RealWorldQA**	**MVBench**	**Video-MME**	**LongVideoBench**
Method Comparison											
Qwen3.5	86.3	84.7	82.5	70.0	46.0	34.7	57.2	75.4	69.1	72.4	67.8
Pure SFT	86.9	82.3	80.6	72.1	47.7	37.5	65.3	75.0	69.5	72.1	68.7
Feat-aligned SFT	87.4	82.3	79.1	72.5	48.0	37.2	65.1	75.6	69.8	72.9	69.4
Stage ablations											
AVTR w/o Stage-0	84.8	80.8	77.1	70.4	47.5	38.2	61.6	75.1	67.1	70.8	67.0
AVTR w/o Stage-1	87.4	86.3	82.5	74.3	53.6	38.4	69.5	80.3	71.8	75.3	68.7
AVTR w/o Stage-3	88.4	86.7	83.8	75.9	50.7	38.4	67.9	76.3	71.5	74.4	69.4
AVTR (default)	89.5	89.4	85.1	75.2	54.2	39.8	71.0	81.5	72.4	76.9	70.5

interface. Stage-1 then makes this visual state question-dependent, while Stage-3 learns when to invoke visual and textual reasoning.

4.5 Comparison With Prior Latent Visual Reasoning

Figure 12 (Appendix) compares both the effectiveness and efficiency of visual imagination over latent visual reasoning methods. The reported Δ measures the gain from enabling visual reasoning over the same method with visual reasoning disabled. Under this paired comparison, our generatively grounded visual imagination yields larger and more consistent improvements across the four high-resolution benchmarks, with an average gain of 1.53 points. Our visual queries are inserted and processed in parallel, avoiding token-by-token latent generation and reducing visual-reasoning latency to 50 ms versus 83-437 ms for prior methods.

4.6 Adaptive Reasoning Improves Performance and Efficiency

Figure 5 shows that visual imagination improves accuracy with only 0.6 seconds of additional latency. This low overhead follows from two design choices. (i) Visual queries are inserted and processed in parallel rather than generated token by token. (ii) The generator is removed at inference, so answers are conditioned directly on the intermediate visual states without pixel decoding.

Adaptive reasoning further selects a route for each question and invokes textual reasoning only when needed. In contrast, the backbone's fixed textual-reasoning mode generates a full reasoning trace for every input. Combining visual and textual reasoning on demand improves the mean score by 4.5 points while reducing latency by 4.3 seconds.

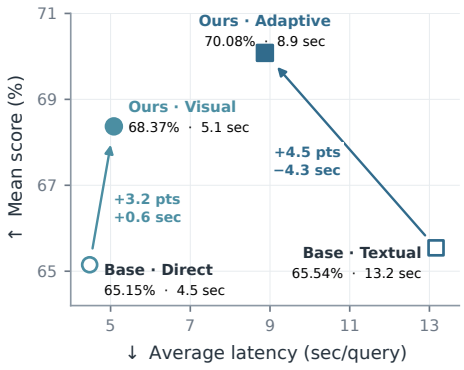

Figure 5: Efficiency analysis.

5 Conclusion

We proposed AVTR, which introduces visual imagination as a complementary reasoning modality and adaptively selects how to reason for each question. Generative visual states revisit task-relevant evidence and directly condition answer generation without pixel decoding at inference. A learned routing policy selects or composes visual imagination and textual reasoning according to the perceptual and inferential demands of each input. Experiments on various image and video understanding benchmarks consistently validate the advantages of our model. We further show that the same framework remains effective with a Qwen3.5-2B backbone, enabling a lightweight adaptive visual-textual reasoning model and highlighting the broader applicability of AVTR across model scales.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, rewriting, and improving the clarity and presentation of the manuscript. We did not use generative AI to generate experimental results or make scientific claims. All AI-assisted text was carefully reviewed, revised, and verified by the authors for technical accuracy, originality, and consistency with the underlying results. We take full responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, et al. Blip3-o: A family of fully open unified multimodal models, architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Xingyu Chen, Jiahao Xu, Tian Liang, et al. Do not think that much for 2+3=? on the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 9487–9499. PMLR, 2025b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024a. doi: 10.1007/s11432-024-4231-5. URL https://doi.org/10.1007/s11432-024-4231-5.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, 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.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations*, volume 2024, pp. 6666–6702, 2024.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: Llm learns when to think. *Advances in Neural Information Processing Systems*, 38:151268–151295, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 24108–24118, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, et al. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

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.

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

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *arXiv preprint arXiv:2602.04476*, 2026.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 36340–36364. PMLR, 2025.

Kunchang Li, Yali Wang, Yinan He, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026b.

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

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, et al. Videoauto-r1: Video auto reasoning via thinking once, answering twice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32829–32839, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025.

OpenAI. Gpt-5 mini model. https://developers.openai.com/api/docs/models/gpt-5-mini, 2025.

OpenBMB. MiniCPM-V 4.6. https://huggingface.co/openbmb/MiniCPM-V-4.6, 2026. Hugging Face model card.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

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.

Qwen Team. Qwen3.5-27B. https://huggingface.co/Qwen/Qwen3.5-27B, 2026a. Hugging Face model card.

Qwen Team. Qwen3.5-2B. https://huggingface.co/Qwen/Qwen3.5-2B, 2026b. Hugging Face model card.

Qwen Team. Qwen3.5-9B. https://huggingface.co/Qwen/Qwen3.5-9B, 2026c. Hugging Face model card.

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. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Xiaoqian Shen and Mohamed Elhoseiny. Visreflect: Latent visual reflection for fine-grained perception in long visual context. In *European Conference on Computer Vision*, pp. 632–649. Springer, 2026.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in. *arXiv preprint arXiv:2512.14273*, 2025.

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.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, et al. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, et al. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026a.

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*, 2025a.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):7907–7915, 2025b.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, et al. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

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*, pp. 13084–13094, 2024.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, 2024. Introduces the RealWorldQA benchmark.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, volume 2025, pp. 20383–20407, 2025a.

Jinheng Xie, Weijia Mao, Zechen Bai, et al. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025b.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184, 2026.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

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

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026a.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33510–33520, 2026b.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, et al. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. Adaptthink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3716–3730. Association for Computational Linguistics, 2025.

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*, 2023.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Chunting Zhou, Lili Yu, Arun Babu, et al. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025a.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

A Training Data and Optimization

The training contains four stages: visual replay, question-conditioned reflection, imagination-guided SFT, and adaptive routing. All phases use bfloat16 precision on NVIDIA A100 80GB GPUs. We report the settings that determine optimization and effective data scale, omitting infrastructure-specific parameters.

A.1 Training Schedule

Stage-0: Visual Replay. Visual replay is trained on approximately 5M captioned images from BLIP3o-Pretrain-Short-Caption (Chen et al., 2025a), CC12M (Changpinyo et al., 2021), and InternVL-SA-1B-Caption (Kirillov et al., 2023; Chen et al., 2024b;a) on 128 A100 GPUs for 50K steps. The Qwen3.5-9B backbone and visual autoencoder are frozen. We optimize the SANA-1.6B generator, 64 visual queries, and 6-layer connector with AdamW (Loshchilov & Hutter, 2019), a peak learning rate of 3×10^{-5}, and cosine decay. All targets are reconstructed at 512×512 resolution. Figures 6 and 7 show representative reconstructions, while Table 8 quantifies the effect of the query budget.

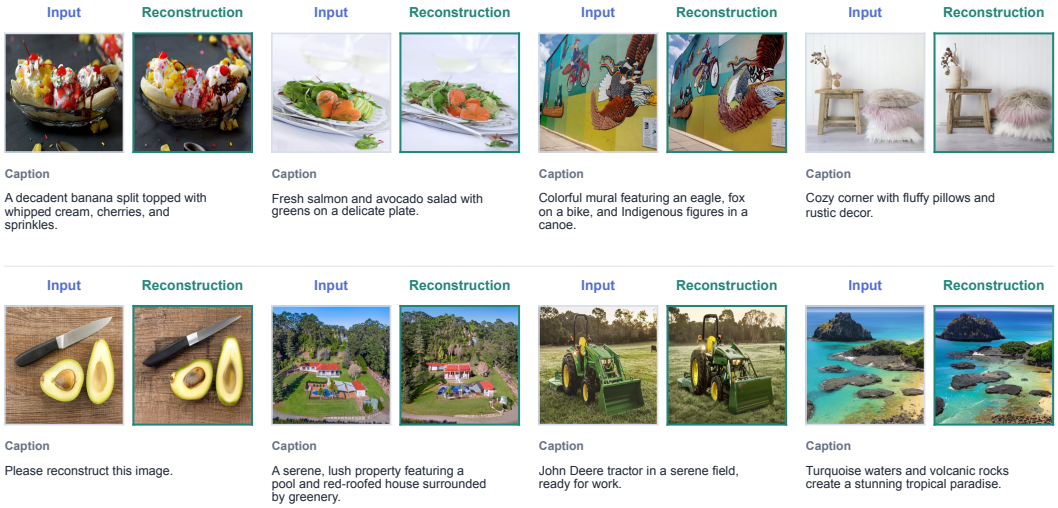

Figure 6: **Visual Reconstruction Quality.** Each example shows an input image, its associated image caption and its reconstruction image.

Stage-1: Visual Reflection. Starting from visual replay, we train on approximately 178K question-conditioned targets retained by the pipeline in Appendix C. The targets comprise salient regions for images and salient frames for videos. We retain the same frozen backbone and autoencoder and update only the generator, queries, and connector. Training uses AdamW with learning rate 3×10^{-5} for one epoch with 64 A100 GPUs.

Stage-2: Imagination-Guided SFT. The SFT mixture contains approximately 380K examples from the sources summarized in Table 6. It combines spatial question answering, general video question answering, and temporal grounding. Target-free examples remain in the mixture and contribute only the response loss. We freeze the generator and visual autoencoder and optimize the VLM, visual queries, and connector for one epoch. We use AdamW with learning rate 10^{-6}, weight decay 0.1, a global batch size of 128, and cosine decay with a 3% warmup with 64 A100 GPUs.

For datasets that also serve as evaluation benchmarks, we train only on their official training splits and evaluate on held-out splits.

Stage-3: Adaptive Routing Data and Optimization. We construct the routing dataset from the SFT training pool. We first perform route-complete rollouts for each question and discard samples

Table 6: **Training data.** The imagination-guided SFT mixture contains approximately 380K image, spatial, and video examples, including 178K data with verified visual target.

Type	Size	Details
Image	28.2K	Visual-CoT (Shao et al., 2024a)
Spatial	191.2K	JoyAI-Image-OpenSpatial (Song et al., 2026)
Video	160.5K	Open-o3-Video (Meng et al., 2025), VideoChat3-LV116K (Li et al., 2026b) ActivityNet Captions (Krishna et al., 2017), Charades-STA (Gao et al., 2017) QVHighlights (Lei et al., 2021), NExT-GQA (Xiao et al., 2024)

for which either all four routes are correct or all four are incorrect, as these cases provide no discriminative supervision for route selection. Among questions solved by exactly one route, we balance the four corresponding route categories so that no reasoning mode dominates the routing data. This results in a training set of 20K questions.

During routing optimization, we force all four routes and sample four rollouts per route for each question. We train with a training batch size of 8, a learning rate of 4×10^{-6}, and GRPO clipping coefficients $\epsilon_l = \epsilon_h = 0.2$. We set the route costs to $C(\textsc{Direct}) = 0$, $C(\textsc{Visual}) = 0.01$, $C(\textsc{Textual}) = 0.04$, and $C(\textsc{Visual+Textual}) = 0.05$. We set $\lambda_{\text{gate}} = 0.02$ based on the approximate gradient magnitudes of the two objectives, keeping the weighted gate gradient on the same order of magnitude as the GRPO gradient.

B Evaluation Protocol

Baselines. The proprietary baselines are GPT-5 (Singh et al., 2025), GPT-5 mini (OpenAI, 2025), and Gemini 2.5 Pro (Comanici et al., 2025). The open foundation models are Qwen3.5-9B and Qwen3.5-2B (Qwen Team, 2026c;b), InternVL3.5-8B and InternVL3.5-2B (Wang et al., 2025a), Cambrian-S-7B (Yang et al., 2025b), LLaVA-OneVision-2-8B (An et al., 2026), JoyAI-Image-Und (Song et al., 2026), Gemma 4 E2B (Gemma Team et al., 2026), and MiniCPM-V4.6 (OpenBMB, 2026). The video-specialized baselines are Video-R1 (Feng et al., 2025), Time-R1 (Wang et al., 2026b), VITAL (Zhang et al., 2026), VideoChat-R1.5 (Yan et al., 2026), and VideoAuto-R1 (Liu et al., 2026). Pure SFT and feature-aligned SFT are controlled variants trained with the same backbone and instruction data. The latent-reasoning comparison includes LVR (Li et al., 2026a), Monet (Wang et al., 2026a), and VisReflect (Shen & Elhoseiny, 2026). The Qwen3.5 rows marked w/o Think and w/ Think use direct answering and the model's built-in textual-reasoning mode, respectively.

Evaluation. All models are evaluated using the same prompts and greedy decoding. Videos are sampled at 2 fps, with at most 256 frames and 20K visual tokens. We set a maximum generation budget of 8,192 new tokens to handle rare cases where textual reasoning fails to terminate. If the model reaches this limit without producing a valid final answer, we fall back to the prediction from the same backbone with thinking disabled, rather than scoring an incomplete reasoning trace.

C Visual Target Construction and Filtering

We construct the reflection targets $\mathcal{V}_{\text{tgt}}$ through candidate localization, target verification, and gain-based filtering. The raw pool contains approximately 0.92M image and video question-answer pairs from full set of data in Table 6. We infer missing targets through model-assisted localization when possible and then verify all targets, yielding approximately 534K examples: 449K images and 84K salient video frames. The final gain-based filter retains approximately 178K examples for reflection training, comprising approximately 155K image targets and 23.5K video targets. These statistics describe the dedicated question-conditioned reflection set.

Candidate Localization. For image examples with a ground-truth bounding box, we directly use it to crop image as the target. When a bounding box is unavailable, Qwen3.5-27B (Qwen Team,

2026a) predicts the region that supports the reference answer. For video examples with an annotated temporal span, we sample a candidate frame from that span. When a temporal span is unavailable, Qwen3.5-27B first localizes the relevant interval, from which we sample the candidate frame. The resulting cropped image or salient frame defines $\mathcal{V}_{\mathrm{tgt}}$.

Target Verification. We next ask Qwen3.5-27B to verify each proposed localization against the original visual input, question, and reference answer. A target passes verification only when the marked region or sampled frame contains visual evidence that directly supports the reference answer. This stage removes empty, irrelevant, and incorrectly localized targets before they are used for training.

Gain-Based Filtering. Verification establishes that a target is relevant, but not that it improves the model's answer. We therefore evaluate the same base VLM twice. The base model receives the original visual input and question, $[\mathcal{V}; q]$, while the target-conditioned run appends the verified target, $[\mathcal{V}; q; \mathcal{V}_{\mathrm{tgt}}]$. For multiple-choice and exact-match questions, we retain wrong-to-correct transitions. For temporal grounding, we require a positive increase in temporal IoU. This paired criterion ensures that each retained target contributes measurable evidence beyond the original input.

D Ablation Studies

Adaptive routing behavior. Table 7 summarizes the routing decisions made by AVTR across benchmarks.

Table 7: **Gate activation under adaptive routing.** Percentage of test examples for which the learned policy activates the visual or textual reasoning gate. The two rates need not sum to 100%, since both gates can be activated for the same example or both can remain inactive.

Benchmark	Visual Gate ON (%)	Textual Gate ON (%)
V*	22.9	34.4
HR-Bench-4K	31.0	77.1
HR-Bench-8K	25.4	74.6
BLINK	52.0	9.8
ERQA	42.2	46.0
MMSI-Bench	30.4	11.9
VSI-Bench	74.3	10.8
RealWorldQA	51.6	71.8
MVBench	66.0	45.6
Video-MME	83.9	52.8
LongVideoBench	66.4	89.7
MLVU	89.0	49.0

Table 8: **Effect of the visual-query budget on reconstruction.** Reconstruction fidelity measured by SSIM (Wang et al., 2004) improves consistently as the number of continuous visual queries increases.

Visual queries (N)	8	16	32	64	128
SSIM	0.2563	0.3934	0.4217	0.5235	0.5632

Effect of the number of visual queries. Table 8 and Figure 7 show that reconstruction quality improves with the number of visual queries. SSIM increases from 0.2563 with 8 queries to 0.5235 with 64 queries and 0.5632 with 128 queries, while 64 queries already produce visually faithful reconstructions. We therefore use 64 queries to balance reconstruction quality and training cost, as larger query sets increase the number of trainable d-dimensional query embeddings.



Figure 7: **Reconstruction quality across visual-query budgets.** Each row shows the input followed by reconstructions with $N = 16$, 32, and 64 queries.

E Qualitative Analysis

E.1 Visual Reflection Visualization

Figures 4 and 8 visualize the evidence retained by the continuous visual queries for video understanding and temporal grounding. The rendered images expose the learned visual state for inspection, while inference uses the continuous queries directly.

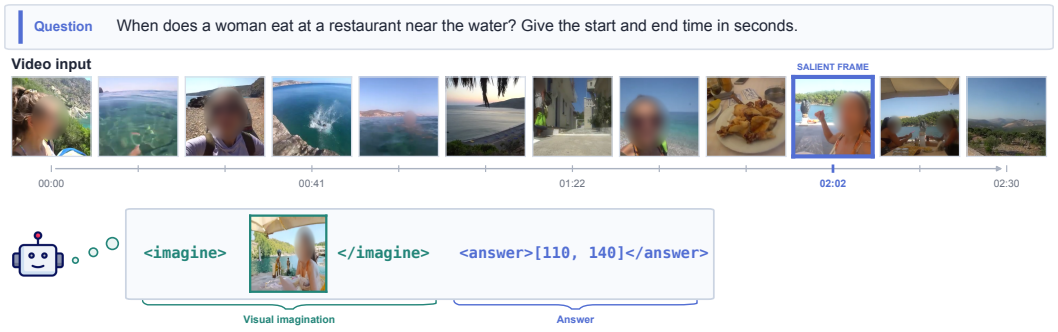

Figure 8: **Visual reflection for temporal grounding.** The rendered image is used only for visualization. The answer is conditioned directly on the continuous visual queries at inference.

E.2 Complementary strengths of visual and textual reasoning

Task-level comparison. We measure route preference as the accuracy difference between the forced visual and textual routes of the same checkpoint. Figure 3 shows a structured continuum rather than a uniform advantage for either modality. The seven largest visual margins all involve metric geometry or viewpoint transformation, spanning room size, multi-view reasoning, relative direction, positional relations, object distance, and object size. Art style and visual similarity show smaller but consistent visual advantages. The strongest textual margins occur in scene sequencing, text-referred tracking, high-resolution cross-instance reasoning, moving direction, and spatial-relation tasks. Action localization, trajectory reasoning, egocentric navigation, counting, functional correspondence, state estimation, and action counting also favor textual reasoning more moderately.

Overall, visual imagination is most useful when the answer depends on reconstructing continuous geometry, whereas textual reasoning better preserves discrete entities, counts, and ordered state changes.

Qualitative examples. Figures 9, 10 and 11 illustrate this complementarity. Visual imagination resolves the viewpoint change in the multi-view example, while textual reasoning better preserves the discrete count in the counting example.

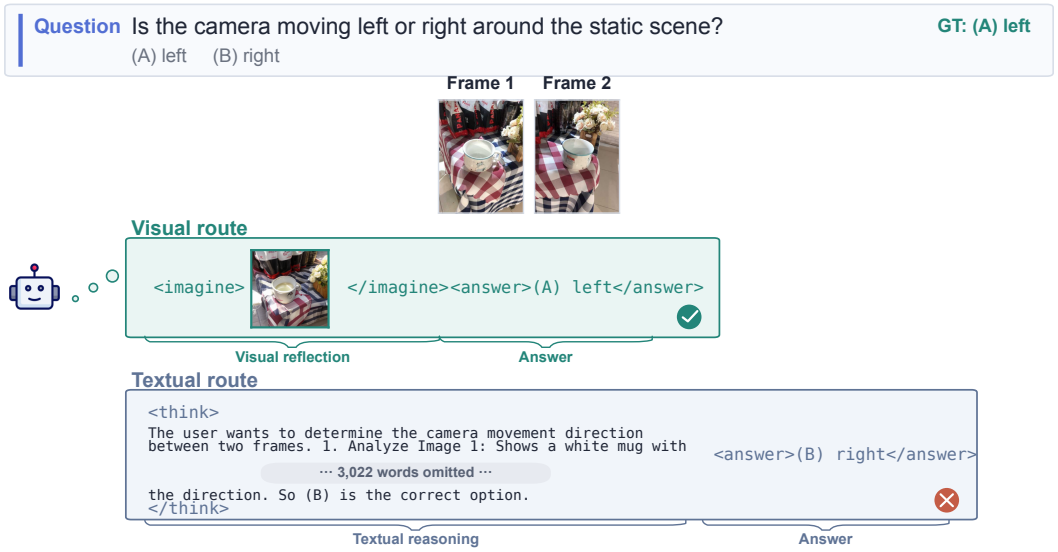

Figure 9: **Example that visual route wins.** On BLINK Multi-view Reasoning, visual imagination produces a useful intermediate view and answers correctly, while the long textual trace reaches the wrong choice.

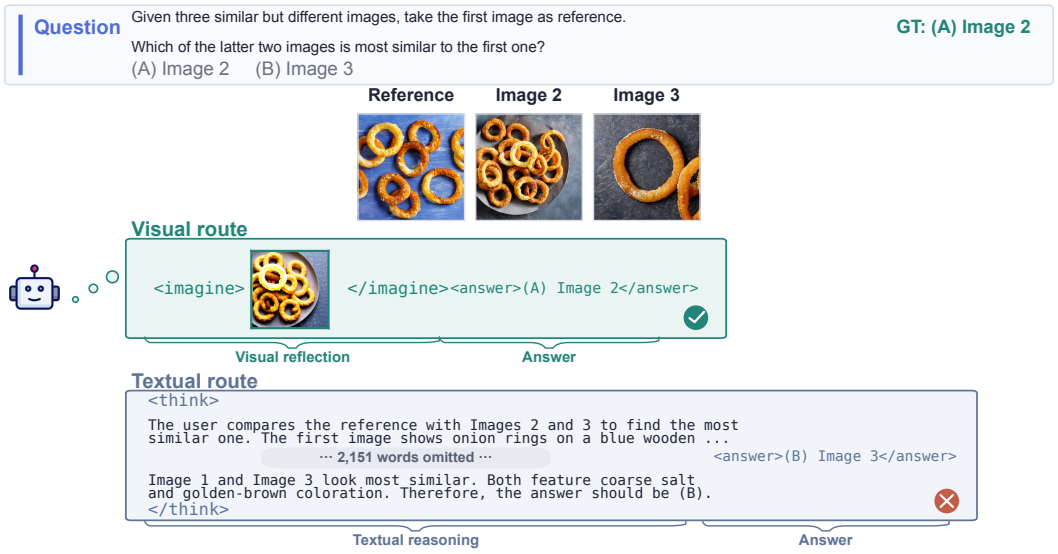

Figure 10: **Example that visual route wins.** On BLINK Multi-view Reasoning, visual imagination produces a useful intermediate view and answers correctly, while the long textual trace reaches the wrong choice.

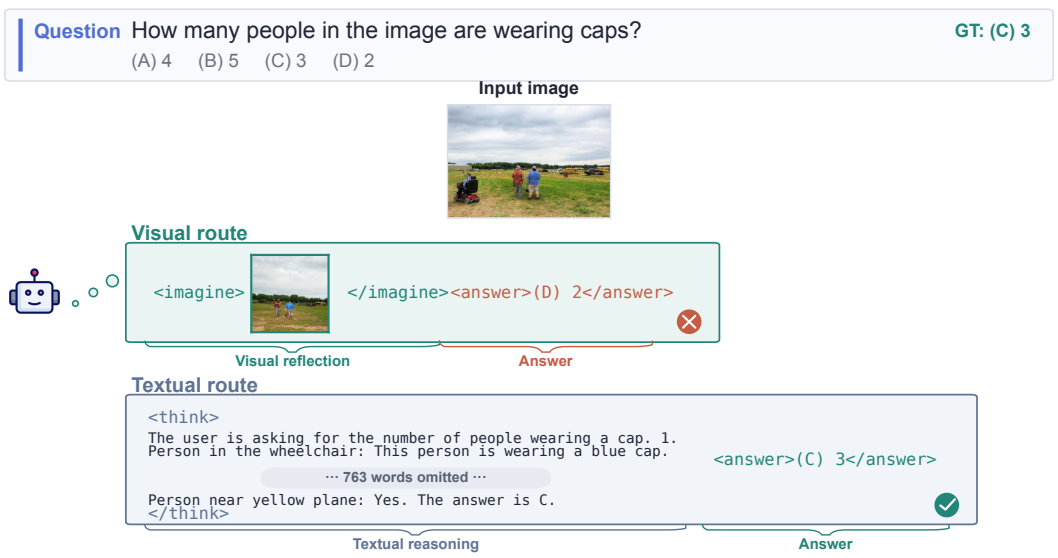

Figure 11: **Example that textual route wins.** On BLINK Counting, textual reasoning returns the correct count, whereas visual imagination changes count-relevant details and produces an incorrect answer.

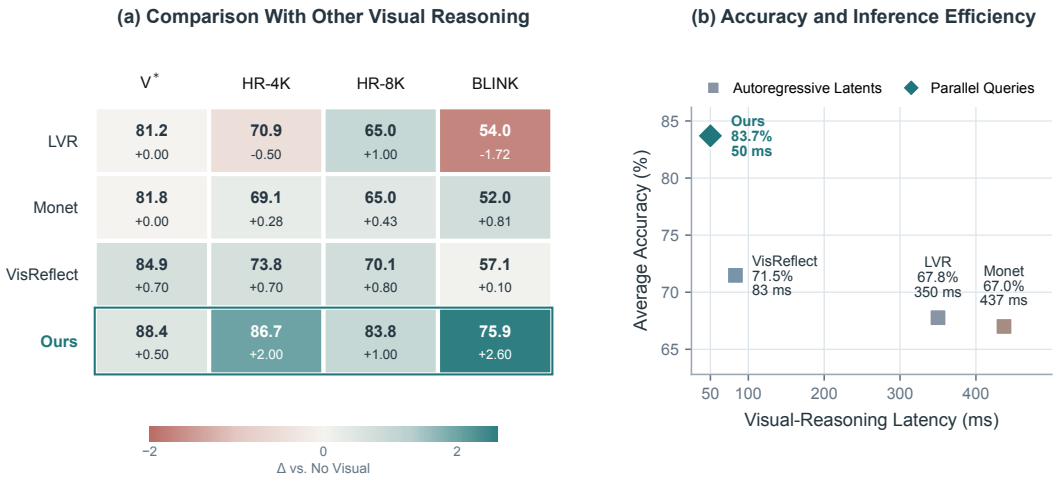

Figure 12: **Generative visual imagination improves both accuracy and efficiency.** Left: accuracy on four high-resolution benchmarks; numbers indicate gains from enabling visual reasoning. Right: average accuracy versus latency.

F Comparison With Prior Latent Visual Reasoning

Prior latent-visual methods differ in both how their states are computed and how those states are supervised. Figure 12(a) reports a paired with-versus-without-visual comparison within each method. For AVTR, these settings correspond to the Direct and fixed Visual routes in Table 4. Enabling our visual state improves all four shared benchmarks, by 1.53 points on average. The corresponding mean changes are -0.31, $+0.38$, and $+0.58$ for Latent Visual Reasoning (LVR), Monet, and VisReflect. For AVTR, the largest gains occur on HR-Bench-4K (+2.0) and BLINK (+2.6), matching the high-resolution setting for which the state is designed.

Figure 12(b) places these results against the reported cost of constructing the visual state. AVTR reaches 83.7 average accuracy at 50 ms, compared with 71.5 at 83 ms and approximately 67 to 68

at 80 to 437 ms for the autoregressive methods. The latency pattern is consistent with the expected advantage of parallel queries over autoregressive latent generation.