Routing by Reasoning Need: Trajectory-Aware Decoding Control for Diffusion Vision-Language Models

Yixiang Liu[1]*, **Zhongxing Xu**[2]*, **Zhonghua Wang**[2]*, **Xiaoying Tang**[1]†,
[1]Southern University of Science and Technology, [2]Monash University
Correspondence: tangxy@sustech.edu.cn

Abstract

Diffusion vision-language models generate answers through iterative refinement, exposing intermediate answer trajectories that can be inspected and controlled at inference time. However, this controllability creates a reasoning-need mismatch, where a universal generation length is applied to questions with different reasoning demands. Visually closed questions may be harmed by continued refinement after a stable answer has formed, whereas reasoning-sensitive questions may be harmed by premature commitment. We formulate this problem as reasoning-budget mismatch and study it in LLaDA-V. Rather than choosing a universal generation length, our training-free controller routes each example to early commitment, baseline preservation, or reasoning-supportive decoding using trajectory signals from answer closure, commitment evidence, and representation revision pressure, without using ground-truth answers. Across answer-focused, mixed-reasoning, and CoT-sensitive benchmarks, routed control improves robustness over fixed long decoding, pure short decoding, and single-rule interventions. The gains are not explained by shorter outputs alone. Answer-closed examples often benefit from commitment, whereas CoT-sensitive examples require preserving or supporting intermediate reasoning. Taken together, these results suggest diffusion VLM decoding should route inference-time control by the state suggested by the observed trajectory instead of relying on a universal decoding length.

1 Introduction

Diffusion vision-language models (VLMs) change what can be controlled during decoding. In left-to-right autoregressive generation, the model commits to a token sequence one position at a time. After a token is emitted, later computation can condition on it but cannot directly revise it. Diffusion language models instead generate through iterative denoising or masked refinement, and recent diffusion VLMs extend this design to multimodal instruction following (Nie et al., 2026; You et al., 2026). This makes decoding a sequence of observable revisions over a partially specified answer rather than a single irreversible trace. Because these intermediate answer trajectories are visible at inference time, they can indicate whether a candidate answer has stabilized, whether the trajectory provides sufficient grounding-and-format evidence for commitment, and whether the model is still changing the representation that will support the final answer. For a visually closed VQA example, such as many compact diagnostic questions in MME (Fu et al., 2025), the answer may appear early and remain consis-

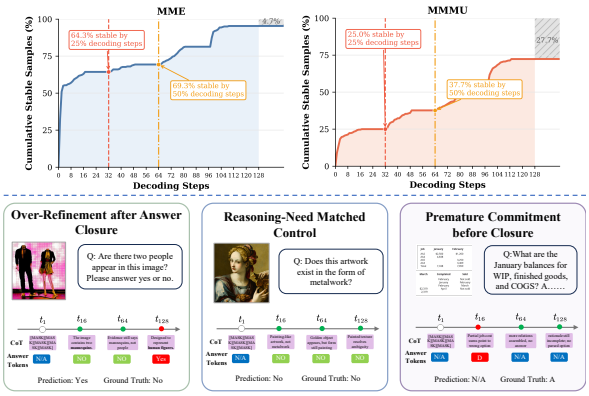


Figure 1: Motivation for reasoning-need matched decoding control. The top row shows answer-closure timing under the fixed 128-step decoding budget across benchmark regimes. Gray regions indicate decoding intervals where no valid answer has converged within the 128-step trajectory. The bottom row illustrates three representative trajectory outcomes, including over-refinement after closure, matched control, and premature commitment before closure. These patterns motivate routing the decoding action by the state suggested by the observed trajectory instead of applying a universal generation length.

* Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21005–21021
October 24-29, 2026 ©2026 Association for Computational Linguistics

tent across refinements (Li et al., 2026b; Xu et al., 2026). By contrast, for geometry, chemistry, or multi-hop reasoning examples, such as MMMU-style expert questions (Yue et al., 2024), the same early surface answer may be unreliable because the model has not yet organized the relevant relations. As shown in the top row of Figure 1, answer closure can occur early for compact visual questions, while reasoning-intensive examples may require later refinement before a reliable answer emerges. Diffusion decoding therefore exposes both the final answer and the revision process that leads to it.

This controllability also creates risk under a fixed decoding budget (Li et al., 2026a). In our experiments, the default fixed-budget reference uses a 128-step refinement trajectory, which assigns every sample the same amount of decoding computation even when questions differ sharply in their reasoning demands. Visually closed questions may be harmed by continued refinement after a stable answer has formed, while reasoning-sensitive questions may be harmed by premature commitment. We call this failure a reasoning-budget mismatch. The mismatch is not simply that some outputs should be shorter and others longer. It is that different samples require different inference-time control actions because the uncertainty left after early refinement has different meanings. In answer-closed cases, the suitable action is early commitment. In unresolved cases, where intervention is not justified, the safer action is to preserve the fixed-budget baseline. In CoT-sensitive cases, where the final answer should remain supported by intermediate reasoning, the suitable action is reasoning-supportive decoding rather than premature commitment.

Figure 1 bottom illustrates this mismatch. Some trajectories reach a stable answer early and are later degraded by additional refinement. Some trajectories remain stable under continued decoding and should be preserved. Others do not yet contain a reliable answer or support-bearing rationale, so early commitment would truncate the reasoning process. These cases suggest that the central question is not whether diffusion decoding should be globally short or long, but which control action is appropriate for the current trajectory.

This paper studies reasoning-budget mismatch in LLaDA-V. Rather than choosing a universal generation length, we propose a training-free inference-time controller that routes each example to one of three actions: early commitment, baseline preservation, or reasoning-supportive decoding. The controller uses trajectory signals computed without ground truth from answer closure, commitment evidence, and representation revision pressure. These signals do not estimate answer correctness or define a ground-truth semantic label of the example. Instead, we use reasoning need as an operational decoding concept: it denotes the control action suggested by the observed diffusion trajectory. This distinction explains why fixed long decoding, pure short decoding, and single-rule interventions can each help in some regimes but fail in others. Diffusion VLM decoding should route inference-time control by the state suggested by the observed trajectory rather than relying on a universal decoding length.

Our contributions are threefold.

- We identify reasoning-budget mismatch in diffusion VLM decoding, where the same refinement budget can over-refine answer-closed examples while under-supporting reasoning-sensitive examples.
- We propose a training-free routed decoding controller for LLaDA-V, which selects among early commitment, baseline preservation, and reasoning-supportive decoding using trajectory signals computed without ground-truth answers.
- We evaluate the controller across answer-focused, mixed-reasoning, and CoT-sensitive settings, showing that routing inference-time control actions is more robust than applying a single global length rule.

2 Related Work

Diffusion VLMs and adaptive decoding. Diffusion language models generate text through iterative denoising or masked refinement rather than left-to-right token commitment (Nie et al., 2026). LLaDA-V extends this formulation to multimodal instruction following and visual reasoning, making intermediate states observable during vision-language generation (You et al., 2026). This process provides a natural interface for inference-time control, since the decoder can inspect partial predictions, uncertainty patterns, and refinement dynamics before the final answer is fixed. Recent adaptive decoding methods exploit such internal signals by adjusting denoising steps, block granularity, token representations, or reasoning modes (Wu et al., 2026; Lu et al., 2026; Yang et al., 2025;

Xu et al., 2026; Li et al., 2026a). These methods show that fixed decoding schedules are often suboptimal, but their objectives are usually efficiency, uncertainty reduction, or hallucination mitigation. Our work studies a different mismatch. The appropriate control action depends on the sample's trajectory state. Some samples are already answer-closed and benefit from early commitment, some remain unresolved, and others require reasoning-supportive decoding rather than shorter decoding.

Reasoning length, grounding, and CoT sensitivity. Longer reasoning is not uniformly beneficial. In multimodal settings, extended reasoning can improve task solving while weakening visual grounding, and reasoning-induced hallucination should be distinguished from perception-induced errors (Xu et al., 2025; Dongre et al., 2025). CoT prompting is task-conditional. It helps when intermediate reasoning supports the decision, but can be unnecessary or harmful when the answer is already recoverable from the input (Sprague et al., 2025; Cheng et al., 2025; Balasubramanian et al., 2025). This suggests that reasoning should be instance-conditional rather than globally enabled or globally suppressed. The relevant decision is not whether to use more or less reasoning for all samples, but whether the current trajectory still requires structured reasoning support. Grounded reasoning methods show that when reasoning is necessary, the better remedy is often to anchor intermediate steps in visual evidence rather than truncate the chain (Man et al., 2025). Our distinction between answer-closed and CoT-sensitive cases follows this view. Early commitment is useful when further refinement mainly adds drift, while preserved or structured reasoning is useful when the decision still depends on unresolved visual-textual evidence.

Decoding-time control and its boundary. Training-free decoding-time methods modify generation without retraining. Representative approaches include visual contrastive decoding, over-trust penalties, layer-contrastive decoding, self-corrective decoding, and memory-based visual refinement (Leng et al., 2024; Huang et al., 2024; Chuang et al., 2024; Huo et al., 2025; Zou et al., 2025). Entropy-aware decoding further shows that token-level uncertainty can guide mode switching and visual-anchor injection during uncertain reasoning states (Xu et al., 2025). These methods demonstrate that internal generation signals can support effective test-time intervention. Our results suggest a boundary for single-rule interventions. Pure short decoding, fixed long decoding, and revision-based control each help in some regimes and fail in others. We therefore treat internal signals not as universal hallucination or factuality detectors, but as evidence for choosing among control actions. This reframes diffusion VLM decoding as a regime-conditioned decision problem, where the system must decide whether to commit early, preserve the baseline path, or apply reasoning-supportive decoding.

3 Method

Figure 2 gives an overview of our reasoning-need routed decoding framework. The controller observes a fixed-budget diffusion trajectory and extracts three signals from intermediate refinement states, including answer closure, grounding-and-format evidence from format and visual-engagement guards, and representation revision pressure. These signals do not use gold labels or estimate correctness directly. They operationalize the state suggested by the observed trajectory as the safest inference-time action under the observed trajectory, routing each example to early commitment, baseline preservation, or reasoning-supportive decoding instead of using a universal short or long budget.

3.1 Problem Setup

We study multimodal question answering with a frozen diffusion VLM. Each input is $x = (x_v, x_q, c, I)$, where x_v is the image, x_q is the question, c denotes optional answer choices, and I denotes an optional reasoning instruction. The model returns an answer $\hat{a}$ and, when requested, a rationale $\hat{r}$.

Let $B = (T, L, M)$ denote the decoding budget, where T is the number of denoising steps, L is the generation length, and M is the block length. The fixed-budget reference uses the same budget for every sample. At denoising step t, the diffusion state is $z_t = (y_t, H_t, P_t)$, where y_t is the token state, H_t denotes hidden states, and P_t denotes token distributions. We write the fixed-budget trajectory and parsed output as

$$\begin{aligned} z_{t+1} &= F_\theta(z_t, x, B), \\ \tau_B(x) &= (z_0, \ldots, z_T), \\ o_B(x) &= \psi(y_T) = (\hat{a}_B, \hat{r}_B). \end{aligned} \tag{1}$$

Here θ is frozen and ψ is the benchmark parser. Our controller does not update model weights, train a

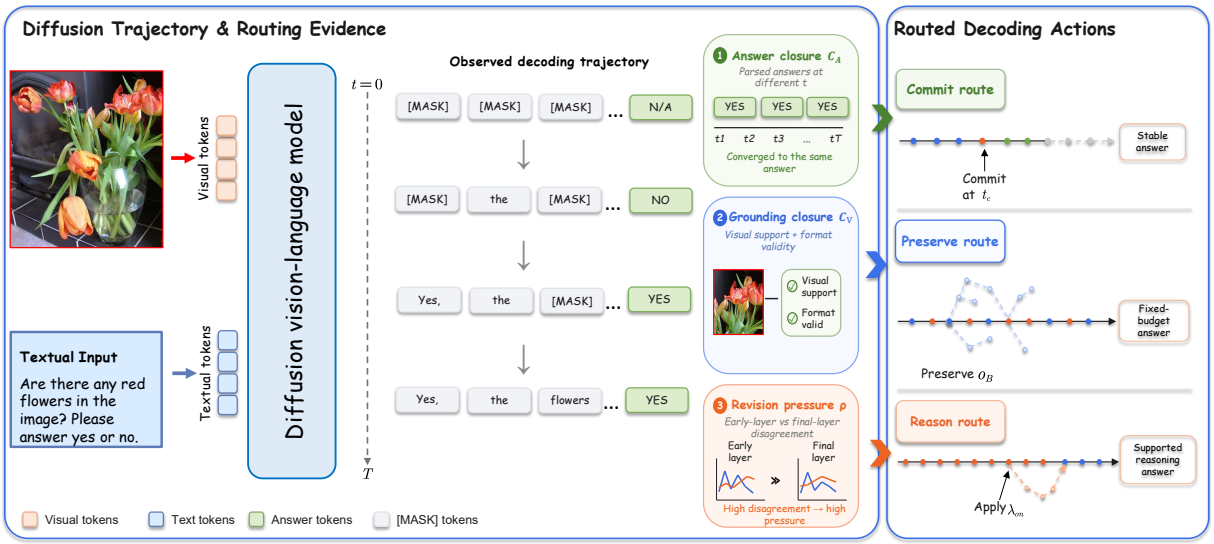


Figure 2: Overview of reasoning-need routed decoding control.

verifier, or use gold labels. It only selects an action from the decoding trajectory.

3.2 Representative Trajectory Patterns

Figure 3 illustrates three answer-token trajectories that motivate our routing policy. The plots show when output positions are decoded, where top-1 predictions change, and whether the answer token remains correct or drifts. The examples correspond to three decoding states: early answer closure with later drift, late convergence that should be preserved, and reasoning-heavy instability that requires additional support. These patterns motivate selecting among early commitment, baseline preservation, and reasoning-supportive decoding according to the sample's reasoning need, rather than using one fixed decoding length.

3.3 Routing Evidence

The router uses three trajectory signals: answer closure, commitment evidence, and representation revision pressure.

Answer closure. Answer closure measures whether the parsed answer has stabilized. Let $\psi_A(y_t)$ return the parsed answer from state y_t, or $\perp$ if no valid answer is available. For a routing window $\mathcal{W}$, let $\hat{a}_{\mathcal{W}}$ be the most frequent valid parsed answer in that window, with $\hat{a}_{\mathcal{W}} = \perp$ if no valid answer appears. We define

$$C_A(x) = \mathbb{1}[\hat{a}_{\mathcal{W}} \neq \perp] \frac{1}{|\mathcal{W}|} \sum_{t \in \mathcal{W}} \mathbb{1}[\psi_A(y_t) = \hat{a}_{\mathcal{W}}]. \tag{2}$$

High C_A means that the same parseable answer repeatedly appears; it does not imply correctness.

visual-format closure. We use visual engagement as a proxy for whether the trajectory attends to image tokens, not whether the answer is true. The proxy aggregates parser validity, visual-attention evidence, prefix-dominance risk, and parsed-answer flip risk:

$$\begin{aligned} \mathbf{s}_V(x) &= \Big(C_{\text{fmt}}(x), C_{\text{vis}}(x), \\ &\qquad R_{\text{prefix}}(x), R_{\text{flip}}(x)\Big), \\ C_V(x) &= g_V(\mathbf{s}_V(x)) \in [0,1]. \end{aligned} \tag{3}$$

Here C_{fmt} checks answer-format validity, C_{vis} is computed from visual-attention mass over image-token positions, and R_{prefix} and R_{flip} capture prefix-dominance and answer-instability risks. The fixed aggregation function g_V provides only routing evidence and does not verify whether the answer is visually true.

Representation revision pressure. Representation revision pressure measures whether early and final layer distributions still disagree. Let $p_{e,t,j}$ and $p_{f,t,j}$ be token distributions from an early layer and the final layer at position j. For route positions Ω_t, we compute

$$\begin{aligned} J_t &= \frac{1}{|\Omega_t|} \sum_{j \in \Omega_t} \text{JS}(p_{e,t,j} \,\|\, p_{f,t,j}), \\ \rho(x) &= \text{Agg}_{t \in \mathcal{W}} J_t. \end{aligned} \tag{4}$$

High $\rho(x)$ indicates that the model is still revising its internal distribution. It is a routing signal, not a hallucination detector or correctness score.

3.4 Routing Policy

The controller chooses among three actions, including CommitRoute, PreserveRoute, and ReasonRoute. Let $K(x) \in \{0, 1\}$ be the fixed CoT-sensitive guard, set from the prompt, task format, or disclosed route configuration. Let $G_{\text{on}}(x) \in \{0, 1\}$ be the fixed route-on guard for the reported operating point. With thresholds fixed before evaluation, define

$$\begin{aligned} U_R(x) &\equiv \rho(x) \geq \beta_\rho \vee C_A(x) < \beta_A, \\ U_E(x) &\equiv C_A(x) \geq \alpha_A \wedge C_V(x) \geq \alpha_V \\ &\quad \wedge \rho(x) \leq \alpha_\rho. \end{aligned} \tag{5}$$

We use $\beta_A \leq \alpha_A$ and $\alpha_\rho \leq \beta_\rho$, so intermediate cases fall back to preservation. The reason and commit conditions are as follows.

$$\begin{aligned} \mathcal{R}(x) &= K(x)\, G_{\text{on}}(x)\, \mathbf{1}[U_R(x)], \\ \mathcal{E}(x) &= \mathbf{1}[U_E(x)]\, \chi_{\text{commit}}(x). \end{aligned} \tag{6}$$

Here $\chi_{\text{commit}}(x)$ denotes fixed hard guards, such as parser availability, operating-point constraints, or route-specific exclusion rules. Signals already summarized by $C_V(x)$ are not counted again in χ_{commit}.

The routing policy is

$$\Pi(x) = \begin{cases} \mathsf{reason}, & \mathcal{R}(x) = 1, \\ \mathsf{commit}, & \mathcal{R}(x) = 0 \wedge \mathcal{E}(x) = 1, \\ \mathsf{preserve}, & \text{otherwise}. \end{cases} \tag{7}$$

The reason condition is evaluated before commitment because answer stability alone is insufficient in CoT-sensitive interfaces. When $\mathcal{R}(x) = 1$, the controller avoids early commitment and applies the disclosed reasoning-supportive configuration. When neither reason nor commit is justified, the controller preserves the fixed-budget output.

3.5 Routing Actions

For ReasonRoute, the controller uses a fixed reasoning-supportive decoding configuration. Let $D_\theta(x; B, \lambda)$ denote the parsed output obtained by running the frozen model under budget B and inference-time configuration λ. Then

$$o_r(x) = D_\theta(x; B, \lambda_{\text{on}}), \tag{8}$$

where λ_{on} is fixed and disclosed for the reported operating point.

For CommitRoute, the controller commits to the first locally stable parsed answer when the enabled closure guards fire. If an online commit step is used, let

$$t_c = \min\{t \mid \mathcal{E}_t(x) = 1\}, \tag{9}$$

where $\mathcal{E}_t(x)$ applies the commit predicate in Eq. 6 to the local routing window ending at step t. Let $\hat{a}_c = \psi_A(y_{t_c})$. The committed output is

$$o_c(x) = A_c(\hat{a}_c, \tau_B, o_B), \tag{10}$$

where A_c is the fixed commit implementation for the reported row. In the main online setting, A_c fixes the answer field to the first stable parsed answer. Trace composition or source selection variants are marked as offline analyses.

PreserveRoute returns the fixed-budget output unchanged:

$$o_p(x) = o_B(x). \tag{11}$$

The final output is

$$o(x) = \begin{cases} o_c(x), & \Pi(x) = \mathsf{commit}, \\ o_p(x), & \Pi(x) = \mathsf{preserve}, \\ o_r(x), & \Pi(x) = \mathsf{reason}. \end{cases} \tag{12}$$

3.6 Routed Decoding Algorithm

Algorithm 1 summarizes the routing procedure. The algorithm computes trajectory evidence, applies the route conditions in Eq. 6, and returns the corresponding routed output. The routine $\mathrm{ObtainRoutingTrace}$ denotes the trace used for routing; rows that require cached candidate traces are marked separately as offline analyses and are not used for latency claims.

4 Experiments

4.1 Experimental Setup

We evaluate LLaDA-V under answer-focused, mixed-reasoning, and CoT or support-sensitive settings. MME (Fu et al., 2025), MMMU (Yue et al., 2024), and MMStar (Chen et al., 2024) use compact final-answer interfaces, but differ in reasoning demands. MME contains many compact visual diagnostic questions, while MMMU and MMStar include more expert-level, fine-grained, or vision-indispensable cases. ScienceQA-IMG

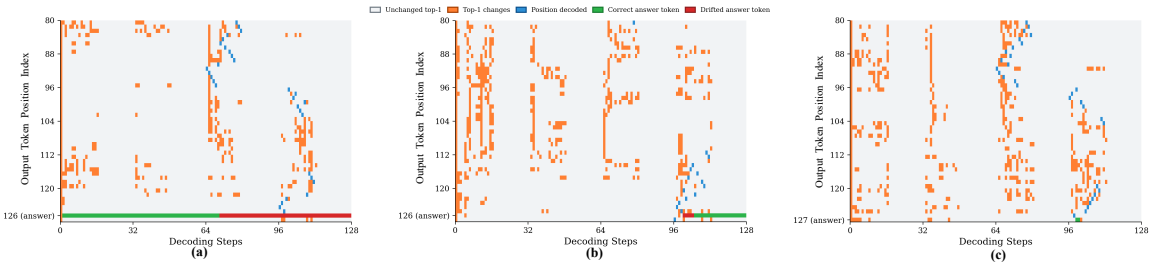


Figure 3: Representative answer-token trajectory patterns that motivate reasoning-need routing. Each panel shows top-1 token changes, decoded positions, and answer-token correctness across diffusion decoding steps. (a) An early-closure case where the correct answer appears early but later refinement drifts to an incorrect answer, motivating early commitment. (b) A late-converging case where the answer becomes reliable only after sufficient refinement, motivating baseline preservation when early commitment is not justified. (c) A reasoning-heavy unstable case where the answer trajectory remains unsettled under simple length control, motivating reasoning-supportive decoding rather than a universal short or long budget.

Algorithm 1 Reasoning-Need Routed Decoding Control

Require: input $x = (x_v, x_q, c, I)$, budget B, frozen model D_θ
Ensure: routed output o

$\tau_B, o_B \leftarrow \text{ObtainRoutingTrace}_\theta(x, B)$
$C_A \leftarrow \text{AnswerClosure}(\tau_B)$
$C_V \leftarrow \text{GroundingFormatClosure}(\tau_B, x)$
$\rho \leftarrow \text{RevisionPressure}(\tau_B)$
$K \leftarrow \text{CoTSensitiveGuard}(x)$
$G_{\text{on}} \leftarrow \text{RouteOnGuard}(x, \tau_B)$
$\chi_{\text{commit}} \leftarrow \text{CommitGuard}(x, \tau_B)$
Compute $U_R, U_E, \mathcal{R}, \mathcal{E}$ using Eqs. 5–6
if $\mathcal{R}(x) = 1$ **then**
    return $D_\theta(x; B, \lambda_{\text{on}})$
else if $\mathcal{E}(x) = 1$ **then**
    $t_c \leftarrow \text{FirstCommitStep}(\tau_B)$
    $\hat{a}_c \leftarrow \psi_A(y_{t_c})$
    return $A_c(\hat{a}_c, \tau_B, o_B)$
else
    return o_B
end if

(Lu et al., 2022), A-OKVQA (An et al., 2024), and MME-CoT (Jiang et al., 2025) are used as support-sensitive settings. ScienceQA-IMG provides annotated explanations, A-OKVQA requires image-grounded commonsense and world knowledge, and MME-CoT directly evaluates multimodal chain-of-thought quality.

All methods use the same frozen LLaDA-V model and the original benchmark parser. We compare fixed-budget baselines with different generation budgets, including Len2, Len32, Len64, and Len128, where Len128 is the default fixed-budget reference. We also compare visual and contrastive decoding controls, adaptive-budget controls, and our routed controller when benchmark-matched artifacts are available. Final-answer accuracy is the primary metric for benchmark comparison. For auxiliary analysis, we report decoding budget statistics in Appendix and CoT rationale diagnostics on the ScienceQA-IMG active-control subset. These diagnostics use reference-based text metrics and GPT-5.5 judge scores, but are not treated as gold-standard human faithfulness evidence.

4.2 Controlled Results Across Regimes

Table 1 summarizes controlled comparisons across answer-focused and support-sensitive settings. The results support three observations.

First, fixed long decoding is not consistently optimal in answer-focused settings. On MME, routed control reaches 79.96, improving over Len128 by 6.55 points and slightly exceeding Len2 at 79.56. On MMMU, it reaches 49.44, improving over Len128 by 7.00 points and exceeding Len2 at 48.89. These gains suggest that routed control is not simply applying a global short budget. Instead, it commits when the trajectory is low-risk and preserves the baseline when early commitment is not justified. On MMStar, routed control reaches 50.77, slightly above Len128 at 49.90, which further supports sample-wise action selection in compact-answer multimodal settings.

Second, final-answer closure can appear even in rationale-bearing benchmarks, but this should not be interpreted as absence of reasoning. On ScienceQA-IMG, routed control reaches 88.60, improving over Len128 by 13.29 points and over Len2 by 4.32 points. Because ScienceQA-IMG provides explanations but Table 1 reports final-answer accuracy, this result shows that final-answer scoring can exhibit early answer closure within a CoT-capable benchmark. It does not imply that the underlying task lacks reasoning structure.

Method	Answer-Focused Evaluation			CoT-Sensitive Evaluation		
	MME	MMMU	MMStar	SQA-IMG	A-OKVQA	MME-CoT
Global budget and task baselines						
Len2	79.56	48.89	46.88	84.28	78.75	48.72
Len32	73.81	28.78	25.00	53.00	82.50	49.57
Len64	74.60	36.11	17.19	76.65	82.50	49.57
Len128	73.41	42.44	49.90	75.31	83.75	49.57
Visual and contrastive controls						
VCD	73.41	47.33	32.81	80.80	83.75	49.29
SID	68.65	47.78	34.38	81.00	82.50	50.43
MEMVR	–	48.56	35.94	–	–	19.37
Adaptive-budget controls						
Fast-dLLM	72.33	42.89	34.38	75.61	83.75	49.57
AdaBlock	71.40	42.33	35.94	74.81	80.00	49.29
DAEDAL-lite	72.03	42.44	32.81	75.90	83.75	49.57
dLLM-Var	69.29	40.89	20.31	70.20	85.00	48.15
Reasoning-need routed control						
Ours	**79.96**	**49.44**	**50.77**	**88.60**	**88.75**	**52.42**

Table 1: Controlled results across answer-focused and CoT-sensitive evaluation settings. Scores are accuracies. Missing entries indicate that no benchmark-matched artifact was available.

Third, pure short decoding is not CoT-safe. On MME-CoT, Len2 falls to 48.72, below Len128 at 49.57, while routed control reaches 52.42, suggesting that premature commitment can remove useful reasoning structure. On targeted A-OKVQA closure, routed control reaches 88.75, outperforming Len128 at 83.75 and the strongest non-routed baseline at 85.00. These patterns support trajectory-aware routing over a universal length policy: short decoding helps stable answers, long decoding preserves reasoning capacity, and routing selects the safer action.

4.3 Length-Policy Stress Test

A central question is whether the improvement comes only from making outputs shorter. The results do not support this explanation. Table 2 compares representative settings where short, long, and routed decoding behave differently. On MME, Len2 is already strong, suggesting that many compact visual questions reach answer closure early. Routed control still slightly improves over Len2 while avoiding a global short policy. On ScienceQA-IMG, routed control improves over both Len2 and Len128, indicating that final-answer gains are not explained by uniformly shortening the output. On MME-CoT, Len2 is worse than Len128, while routed control improves over both.

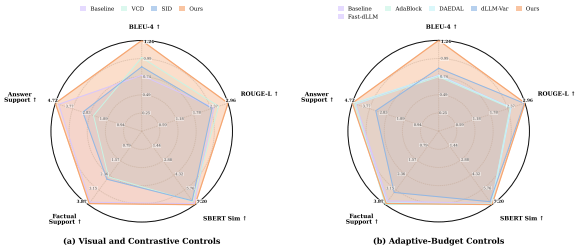


Figure 4: Auxiliary CoT rationale-quality diagnostics on the ScienceQA-IMG active-control subset. The axes include BLEU-4, ROUGE-L, SBERT similarity, factual support, and answer support. The two judge dimensions correspond to Reference-Grounded Factuality and Rationale-Answer Alignment. The left panel compares visual and contrastive controls, while the right panel compares adaptive-budget controls. Both panels use the same zero-based per-axis normalization for visualization. These diagnostics are auxiliary and should not be interpreted as evidence of causal rationale faithfulness.

This shows that early commitment is unsafe when reasoning support is required.

Table 3 gives a complementary route-level diagnostic. If routed control simply behaved like short decoding, most examples would be routed to commitment across all settings. Instead, the allocation changes with the evaluation interface and trajectory state. MME is dominated by commitment, which matches its compact answer-focused interface. In

Setting	Len2	Len128	Ours	Steps
MME	79.56	73.41	**79.96**	128.00
SQA-IMG	84.28	75.31	**88.60**	118.09
MME-CoT	48.72	49.57	**52.42**	102.09

Table 2: Length-policy stress test. Scores are accuracies, and Steps reports the average decoding steps of routed control. The pattern shows that routed control is not equivalent to uniformly shortening the output.

Setting	Commit	Preserve	Reason
MME	75.99	24.01	0.00
SQA-IMG†	0.00	88.75	11.25
MME-CoT	0.00	71.79	28.21

Table 3: Diagnostic route distribution across evaluation settings. Values are percentages. †ScienceQA-IMG uses the CoT diagnostic route interface.

ScienceQA-IMG and MME-CoT diagnostic CoT settings, premature commitment is suppressed, and the controller mainly preserves the baseline or applies reasoning-supportive decoding.

Together, the two diagnostics clarify the role of routing. Short decoding helps when the answer is already closed, but it can remove useful intermediate structure in CoT-sensitive settings. Long decoding preserves reasoning capacity, but it can continue refining samples whose answer has already stabilized. Routed control improves robustness by selecting the safer action under the observed trajectory, rather than by committing to one global length policy.

4.4 Auxiliary CoT Diagnostics on ScienceQA-IMG

Final-answer accuracy does not show whether a generated reasoning trace resembles the reference explanation or supports the predicted answer. We therefore add an auxiliary CoT diagnostic on the ScienceQA-IMG active-control subset. For automatic evaluation, we compare each generated CoT trace with the ScienceQA reference explanation and report BLEU-4, ROUGE-L, and SBERT similarity. BLEU-1 is reported in Appendix as a supplementary overlap metric. For judge-based diagnostics, GPT-5.5 scores two text-only support dimensions on a 0–10 scale, factual support and answer support. The judge sees the question, reference explanation, predicted answer, and generated reasoning trace, but these scores are used only as auxiliary diagnostics.

Figure 4 and Table 6 show that routed control improves reference-based explanation metrics over the fixed long-budget baseline. BLEU-4 increases from 0.105 to 0.124, ROUGE-L from 0.268 to 0.296, and SBERT similarity from 0.697 to 0.720. The paired tests are significant for all three metrics, with $p = 1.03\times10^{-5}$ for BLEU-4, $p = 8.53\times10^{-6}$ for ROUGE-L, and $p = 3.41\times10^{-3}$ for SBERT similarity. The judge scores provide complementary evidence that routed control improves answer support while remaining competitive on factual support.

These results are auxiliary rather than definitive evidence of reasoning faithfulness. They show that routed control can improve actively controlled reasoning traces on a rationale-bearing benchmark, but they do not establish that the generated rationales faithfully explain the model's predictions.

5 Conclusion

Diffusion VLMs expose intermediate answer trajectories before the final output is fixed. This makes fixed-budget decoding a poor fit for questions with different closure times. Some examples already contain a stable, visually supported answer and can be harmed by continued refinement. Others require additional refinement, so early commitment can remove a useful intermediate structure.

We study this problem as reasoning-budget mismatch and address it with a parameter-frozen routed controller. The controller chooses among early commitment, fixed-budget preservation, and reasoning-supportive decoding from trajectory evidence rather than from a global length rule. Across answer-focused, mixed-reasoning, and CoT-sensitive settings, this routing view avoids the main failure modes of both pure short decoding and fixed long decoding. The broader implication is that diffusion VLM decoding should be treated as trajectory-aware control, not as universal generation-length selection.

Limitations

This study has several limitations. First, the controller is evaluated on LLaDA-V, and additional diffusion VLMs are needed to establish model-level generality. Second, the routing signals are label-free proxies rather than correctness estimators. In particular, visual-attention evidence measures image-token engagement, not whether the answer is visually true. Third, some operating

points use benchmark-specific guards or reasoning-supportive configurations; although these are fixed before evaluation, future work should study more unified threshold selection and calibration. Fourth, our CoT diagnostics combine reference-based automatic metrics with judge-based support scores, which should not be interpreted as causal evidence of rationale faithfulness. Finally, while routed control improves robustness, it is not primarily an acceleration method in the reported setting, since some routes retain long decoding budgets to preserve reasoning support.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (T2422012); the National Key Research and Development Program of China (2023YFC2415400); the Guangdong Basic and Applied Basic Research (2024B1515020088); the Shenzhen Science and Technology Program (ZDYJ20251211121037006); the High Level of Special Funds (G030230001, G03034K003); the Guangdong S&T Program (2025B1111080001); the SUSTech Fang Keng Faculty Award.

References

Wenbin An, Feng Tian, Jiahao Nie, Wenkai Shi, Haonan Lin, Yan Chen, QianYing Wang, Yaqiang Wu, Guang Dai, and Ping Chen. 2024. Knowledge acquisition disentanglement for knowledge-based visual question answering with large language models. *arXiv preprint arXiv:2407.15346*.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. 2025. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. In *EMNLP (Findings)*, pages 13406–13439.

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

Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. 2025. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation. *arXiv preprint arXiv:2506.17088*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.

Vardhan Dongre, Chi Gui, Shubham Garg, Hooshang Nayyeri, Gokhan Tur, Dilek Hakkani-Tür, and Vikram S Adve. 2025. Mirage: A benchmark for multimodal information-seeking and reasoning in agricultural expert-guided conversations. *arXiv preprint arXiv:2506.20100*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427. IEEE.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 24272–24295.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. MME-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *International Conference on Machine Learning*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882. IEEE.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. 2026a. Beyond fixed: Training-free variable-length denoising for diffusion large language models. In *International Conference on Learning Representations*, volume 2026, pages 91715–91731.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. 2026b. Diffusion language model knows the answer before it decodes. In *The Fourteenth International Conference on Learning Representations*.

Guanxi Lu, Hao Chen, Yuto Karashima, Zhican Wang, Daichi Fujiki, and Hongxiang Fan. 2026. Adablock-dllm: Semantic-aware diffusion llm inference via adaptive block size. In *International Conference*

on Learning Representations, volume 2026, pages 140018–140033.

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

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14268–14280. IEEE.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2026. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646.

Zayne Sprague, Fangcong Yin, Juan Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, volume 2025, pages 94118–94162.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2026. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. In *International Conference on Learning Representations*, volume 2026, pages 57027–57051.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905.

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

Yicun Yang, Cong Wang, Shaobo Wang, Zichen Wen, Biqing Qi, Hanlin Xu, and Linfeng Zhang. 2025. Diffusion llm with native variable generation lengths: Let [eos] lead the way. *arXiv preprint arXiv:2510.24605*.

Zebin You, Shen Nie, Xiaolu Zhang, JUN ZHOU, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2026. Llada-v: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10093–10105.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. 2025. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Forty-second International Conference on Machine Learning*.

A Appendix

A.1 Available Diagnostic Comparisons

Table 5 reports action-level diagnostic comparisons for settings where corresponding artifacts are available. These rows are not intended as a full component ablation. They compare the full routed setting with available preserve-only, commit-only, separated-source, and guarded-routing variants.

The comparisons show two patterns. First, the full router improves over preserve-only baselines across all reported settings. Second, the full router is not equivalent to a commit-only policy. Commit proxy is competitive on answer-focused settings, but it underperforms the full router and drops on MME-CoT, where premature commitment can remove reasoning support. The ScienceQA-IMG CoT separated-source setting reaches the same answer accuracy as the full routed setting, suggesting that answer accuracy alone does not capture the effect of source selection in this diagnostic protocol. On MME-CoT, the guarded route improves over the unguarded route variant, indicating that the final guard helps reduce harmful route-on cases. We treat these rows as diagnostic action comparisons, while full component isolation is left to future controlled ablations.

A.2 Routing Guards and Operating Points

$K(x)$ is an interface-level CoT-sensitive guard. It is enabled when the benchmark prompt requests a rationale, the evaluation interface scores reasoning support, or a disclosed route configuration marks the input as CoT-sensitive. $G_{\text{on}}(x)$ is the fixed route-on guard for the reported operating point. $\chi_{\text{commit}}(x)$ contains hard exclusions such as parser unavailability, operating-point constraints, or route-specific exclusion rules.

For MME-CoT, λ_{on} corresponds to the answer-first CoT prompt with `jsd_adaptive_throttle` and fixed route-on thresholds. When Route v3 is reported, it should be interpreted as a disclosed MME-CoT operating point rather than a benchmark-independent routing rule. Rows that use cached fixed-budget and route-on candidate outputs are marked as offline trace-composition analyses and are not used for online latency claims.

A.3 Budget and Efficiency Analysis

Table 4 reports the budget controls. Fixed128 / Long128 decoding is computationally expensive across MME, ScienceQA-IMG, MME-CoT, and MMMU. Short2 is much cheaper and is often strong in answer-closed settings, reaching 79.56 on MME, 84.28 on ScienceQA-IMG, and 48.89 on MMMU with a much smaller decoding budget. This confirms that many answer-closed examples do not need the full fixed refinement path.

However, Short2 is not CoT-safe: on MME-CoT it reaches 48.72, below the Fixed128 CoT baseline at 49.57. Intermediate budgets also do not consistently solve the mismatch. Short32 and Fixed64 are below Short2 on MME and ScienceQA-IMG, and they do not improve over Fixed128 on MME-CoT. Route v3's MME-CoT gain is not explained by shorter outputs alone because its average steps and latency are comparable to the long CoT baseline. The method is therefore not simply reducing length; it selectively allocates or structures reasoning. These results motivate selective budget allocation rather than uniform budget reduction.

Benchmark	Method	Steps	Time	Len.
Answer-closed / mixed settings				
MME	Len128	128.00	22.53s	47.0
MME	Len64	64.00	10.60s	23.43
MME	Len32	32.00	4.85s	11.40
MME	Len2	**2.00**	**0.28s**	1.0
MME	Ours	128.00	22.77s	40.01
CoT-sensitive setting				
MME-CoT	Len128	95.91	25.52s	89.42
MME-CoT	Len64	40.05	10.41s	45.05
MME-CoT	Len32	28.09	4.97s	23.55
MME-CoT	Len2	**2.00**	0.35s	**1.46**
MME-CoT	Ours	102.09	25.67s	89.26
ScienceQA-IMG	Len128	101.47	24.6s	99.94
ScienceQA-IMG	Len64	64.00	11.10s	23.57
ScienceQA-IMG	Len32	32.00	5.64s	11.15
ScienceQA-IMG	Len2	**2.00**	0.31s	**1.00**
ScienceQA-IMG	Ours	118.09	24.89s	110.33

Table 4: Budget and efficiency controls. The table reports decoding budget rather than answer accuracy. Entries marked with are approximate estimates from measured budget trends when full audit latency or output length was unavailable. Len2 is inexpensive for answer-closed settings, whereas the MME-CoT route retains a long reasoning budget, indicating that its gain is not an output-shortening effect.

A.4 Shared Decoding Settings

All reported rows use the frozen LLaDA-V model with the benchmark image input and task prompt

Variant	MME	MMMU	SQA-CoT	MME-CoT
Full router	**79.96**	**49.44**	**88.60**	**52.42**
Preserve only	73.41	42.44	84.18	49.57
Commit proxy	79.56	48.89	NA	48.72
Separated source	NA	NA	88.60	NA
Unguarded route	NA	NA	NA	51.57

Table 5: Available diagnostic comparisons for routed decoding. Entries are accuracies. Preserve only denotes the fixed-budget or same-setting joint CoT baseline. Commit proxy uses the shortest benchmark-matched answer proxy. Separated source is reported only for ScienceQA-IMG CoT. Unguarded route is reported only for MME-CoT.

Metric	Baseline	Method	Delta	p
BLEU-1	0.269	0.280	+0.011	1.09×10^{-4}
BLEU-4	0.105	0.124	+0.018	1.03×10^{-5}
ROUGE-L	0.268	0.296	+0.028	8.53×10^{-6}
SBERT sim	0.697	0.720	+0.023	3.41×10^{-3}

Table 6: Automatic ScienceQA-IMG CoT explanation metrics on the active-control subset. BLEU-4, ROUGE-L, and SBERT similarity are used in Figure 4; BLEU-1 is reported as a supplementary overlap metric.

assembled by the evaluation scripts. We do not update model weights or train a verifier. Unless a method-specific control requires a different field, decoding uses deterministic sampling with temperature 0, low-confidence remasking, and the benchmark parser used by the corresponding evaluation.

Fixed-budget rows keep the same model, prompt, and parser while changing the diffusion budget. Len2, Len32, Len64, and Len128 use generation lengths matched to the corresponding number of denoising steps. Adaptive-budget controls keep the same model and parser, while using their method-specific decoding rules.

A.5 Routing Configuration Disclosure

Table 8 summarizes the routing configuration used in each evaluation setting. All routing rules are fixed before evaluation and do not use gold labels. For answer-focused settings, the controller mainly decides whether the trajectory is safe to commit or should be preserved. For CoT-sensitive settings, the reason route can be enabled when the interface requires support-bearing reasoning and the observed trajectory remains unresolved.

Rows that use cached candidate outputs are used only for diagnostic composition and are not used for online latency claims. This preserves the same benchmark prompt, parser, model weights, and branch-level decoding configuration, while separating diagnostic route analysis from online efficiency evaluation.

Field	Setting
Model	Frozen LLaDA-V
Temperature	0.0
Remasking	Low confidence
Default budget	128 steps, 128 generation length
Short budgets	2, 32, and 64 steps
Parser	Original benchmark parser
Trace collection	Enabled for routed settings

Table 7: Shared decoding settings used in the reported experiments.

Setting	Commit	Preserve	Reason
ScienceQA-IMG	enabled	enabled	setting dependent
MME	enabled	enabled	disabled
MMMU	enabled	enabled	enabled
MME-CoT	disabled	enabled	enabled

Table 8: Routing actions enabled in each evaluation setting. ScienceQA-IMG uses a CoT diagnostic route interface for the rationale-quality analysis.

A.6 LLM-Judge Prompt for ScienceQA-IMG CoT Diagnostics

We use a text-only LLM judge as an auxiliary diagnostic for ScienceQA-IMG CoT outputs. The judge does not inspect the image and does not see method names. It receives the question, answer choices, reference explanation, predicted answer, and generated rationale. The scores are used only to assess local support properties of the generated rationale and are not treated as evidence of causal rationale faithfulness.

```
You are evaluating a generated rationale for a ScienceQA-
      IMG multiple-choice example. This is a text-only
      evaluation. You cannot inspect the image. Use only
      the question, answer choices, reference explanation,
       predicted answer, and generated rationale. Do not
      infer unseen image contents. Do not use method names.
       Return JSON only.

Evaluate only the generated rationale. Metrics, each from
      0 to 10:

1. factual_support: Does the rationale make claims that
      are supported by the question, answer choices,
      predicted answer, and reference explanation? A high
      score means the rationale is specific, plausible,
      and does not introduce unsupported facts.

2. answer_support: Does the rationale coherently support
      the predicted answer? A high score means the
      rationale explains why the predicted answer follows,
       rather than drifting toward another option or
      contradicting the predicted answer.
```

```
Scoring guide: 10 = accurate, specific, coherent, and well
       supported. 8 = mostly supported with minor
      omissions or harmless imprecision. 6 = partially
      supported but generic or incomplete. 4 = weak
      support, notable unsupported claims, or loose
      connection. 2 = mostly unsupported, contradictory,
      repetitive, or off-topic. 0 = empty, invalid,
      nonsensical, or supports a different answer.

Question: {question}
Answer choices: {choices}
Reference explanation: {reference_explanation}
Predicted answer: {predicted_answer}
Generated rationale: {generated_rationale}

Return exactly:
{
  "factual_support": integer from 0 to 10,
  "answer_support": integer from 0 to 10,
  "short_reason": "one short sentence"
}
```

Judge calls use deterministic decoding with temperature 0. The judge outputs are parsed as JSON. We cache judge results by example and method to avoid repeated calls. We do not request or store hidden chain-of-thought from the judge.

A.7 Illustrative ScienceQA-IMG non-preserve CoT examples

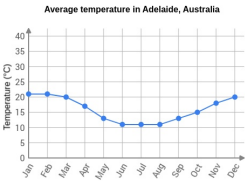
Illustrative ScienceQA-IMG non-preserve CoT examples (Case1–Case3). Each panel compares Len128 with Ours on the same example id and reports automatic explanation metrics and text-only judge scores. These diagnostics support rationale-quality analysis but do not prove faithfulness.

A.8 Illustrative MME answer-token drift cases under fixed 128-step decoding.

Illustrative MME answer-token drift cases (Case4 and Case 5) under fixed 128-step decoding. Green marks the correct answer-token interval, red marks the drifted answer-token interval, blue marks decoded positions, and orange marks top-1 changes. These examples diagnose answer-slot drift and do not constitute additional benchmark results.

scienceqa_full_01985: ScienceQA-IMG CoT Example

Selected from the non-preserve subset where Ours improves over Len128; illustrative rationale-quality diagnostic.



Question

Which month is the hottest on average in Adelaide? A. June, July, and August B. April and May C. January and February

Choices / Gold

A. June, July, and August B. April and May C. January and February

Gold: C (January and February)

Reference explanation

To describe the average temperature trends in Adelaide, look at the graph. Choice "Feb" is incorrect. Choice "Apr" is incorrect. Choice "May" is incorrect. Choice "Jun" is incorrect. Choice "Jul" is incorrect. Choice "Aug" is incorrect. Choice "Jan" is incorrect. The average temperatures in January and February are just above 20°C. These months have the highest average temperatures of all of the months. So, they are hotter than the other months.

Len128 Baseline

Pred: A | correct=False
BLEU-4 3.6 ROUGE-L 21.3 SBERT 70.0
Factual 0 Answer 0 Halluc. safety 0 Avg 5

I will analyze the line graph to determine which month has the highest average temperature in Adelaide. I will identify the peak point on the graph and match it with the corresponding month. The graph displays the average temperature in Adelaide, Australia, from January to December. It shows a general upward trend peaking in June, July, and August. To find the hottest month on average, I will look for the highest point on the graph. The graph shows that the temperature reaches its peak in June, July, and August. This indicates that the average temperature in Adelaide is highest during these months. A ```

Judge note: It incorrectly claims the graph peaks in June-August, contradicting the reference that January and February are hottest.

Ours

Pred: C | correct=True
BLEU-4 6.9 ROUGE-L 23.4 SBERT 81.1
Factual 100 Answer 100 Halluc. safety 100 Avg 95

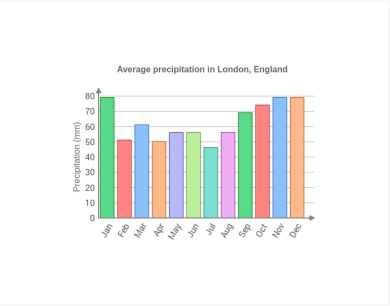
The graph displays the average temperature in Adelaide, Australia, from January to December. The average temperature in Adelaide is highest during the months of January and February, as indicated by the data points on the graph. The graph shows that the average temperature in Adelaide is the months of in January and February.

Judge note: It correctly identifies January and February as the hottest months and stays consistent with the reference, despite minor awkward wording.

Figure 5: Case1

scienceqa_full_03285: ScienceQA-IMG CoT Example

Selected from the non-preserve subset where Ours improves over Len128; illustrative rationale-quality diagnostic.



Question

Which month has the lowest average precipitation in London? A. November B. July C. February

Choices / Gold

A. November B. July C. February

Gold: B (July)

Reference explanation

To describe the average precipitation trends in London, look at the graph. Choice "Feb" is incorrect. Choice "Jul" is incorrect. Choice "Nov" is incorrect. July has an average monthly precipitation of about 45 millimeters. This is lower than in any other month. So, July has the lowest average precipitation.

Len128 Baseline

Pred: C | correct=False

BLEU-4 4.4 ROUGE-L 20.8 SBERT 72.4

Factual 25 Answer 0 Halluc. safety 25 Avg 30

I will examine the bar chart to identify the month with the lowest average precipitation in London by comparing the heights of the bars for each month. The bar chart displays the average precipitation in London, England, from January to December. Each bar represents the average precipitation in millimeters, with different colors representing different months. To find the month with the lowest average precipitation, I will look for the shortest bar on the chart. By comparing the heights of the bars for each month, I can determine which month has the least precipitation. The bar for February is the shortest, indicating the lowest precipitation. C ```

Judge note: It clearly answers February, but that contradicts the reference that July has the lowest precipitation and includes unsupported chart details.

Ours

Pred: B | correct=True

BLEU-4 12.1 ROUGE-L 30.1 SBERT 87.2

Factual 75 Answer 100 Halluc. safety 75 Avg 85

The graph displays the average precipitation in London, England, from January to December. To determine which month has the lowest average precipitation, we need to the heights of the bars each month. . Based on these observations, we can conclude that July is the month with the lowest average precipitation in London.

Judge note: It correctly identifies July as the lowest month and gives a simple graph-based rationale without notable unsupported details.

Figure 6: Case2

scienceqa_full_01410: ScienceQA-IMG CoT Example

Selected from the non-preserve subset where Ours improves over Len128; illustrative rationale-quality diagnostic.



Question

Does this passage describe the weather or the climate? A. climate B. weather

Choices / Gold

A. climate B. weather

Gold: B (weather)

Reference explanation

Read the passage carefully. Madison, the capital of Wisconsin, is located in the southern part of the state. For two weeks in December, the temperature never rose above 20°F. The underlined part of the passage tells you what the temperature in Madison was like in December. This passage describes the atmosphere at a certain place and time. So, this passage describes the weather.

Len128 Baseline

Pred: B | correct=True

BLEU-4 4.3 ROUGE-L 21.7 SBERT 39.1

Factual 25 Answer 50 Halluc. safety 0 Avg 25

To determine whether the passage describes the weather or the climate, I will analyze the visual cues in the image and compare them to the options. The image shows a large building with a dome, surrounded by snow-covered trees and ground. The sky is overcast, suggesting a cold, snowy environment. The presence of snow and the overcast sky indicate a specific condition at a particular time, which is characteristic of weather. Climate, on the other hand, refers to long-term averages and patterns. Since the image depicts a specific moment with snow and overcast skies, it describes the weather. B ```

Judge note: It selects the correct answer, but the explanation invents image details instead of using the passage-based evidence in the reference.

Ours

Pred: B | correct=True

BLEU-4 21.8 ROUGE-L 34.8 SBERT 82.9

Factual 100 Answer 100 Halluc. safety 100 Avg 100

Evidence 1: The passage mentions that the temperature in Madison, Wisconsin, never rose above 20°F for two weeks in December. This is a specific description of the atmosphere at a particular time and place. Evidence 2: The passage describes a short-term event, which is characteristic of weather.

Judge note: It correctly identifies the passage as describing short-term atmospheric conditions at a specific place and time.

Figure 7: Case3

mme_0881: Answer-Token Drift Case

Illustrative appendix example; not an additional benchmark result.



Question

Are there two people appear in this image?
Please answer yes or no.

Gold: No

Fixed128 final: Yes (correct=False)

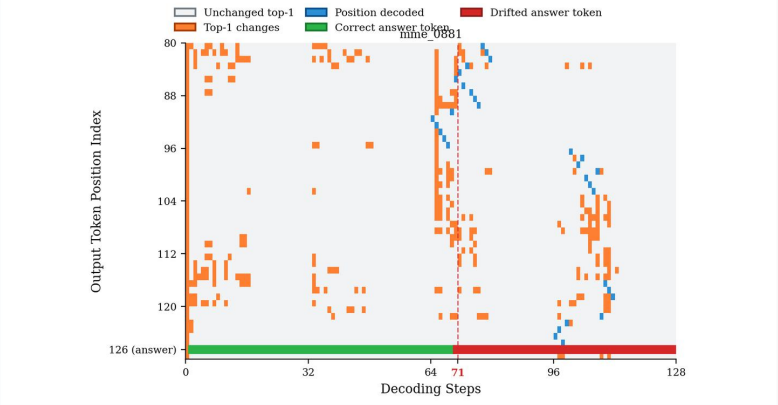
First wrong-after-correct step: 71; stable wrong from step: 71.

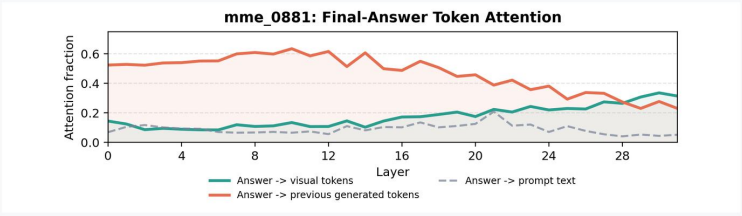
Read the heatmap as answer-slot behavior over diffusion steps: green denotes the correct answer token interval, red denotes the drifted answer-token interval, and blue marks positions decoded at each step.

Diagnostic error mechanism

The trace observes mannequins but then treats the presence of two mannequin figures as support for two people, creating an answer-type mismatch. Answer drift: correct answer persists through step 70, then the final answer stabilizes as Yes from step 71.

Attention diagnostic: Final answer token `Yes` attends more to generated prefix than visual tokens on average (0.468 vs 0.173); late-layer prefix/visual ratio = 1.08x.





Interpretation is diagnostic: the pattern is consistent with self-anchoring on the generated prefix, not a standalone causal proof.

Figure 8: Case4

mme_0061: Answer-Token Drift Case

Illustrative appendix example; not an additional benchmark result.



Question

Is this artwork created by morel, jean-baptiste?
Please answer yes or no.

Gold: No

Fixed128 final: Yes (correct=False)

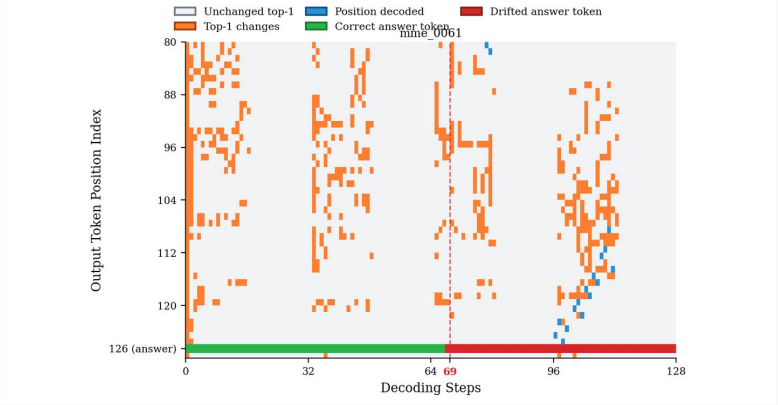
First wrong-after-correct step: 69; stable wrong from step: 69.

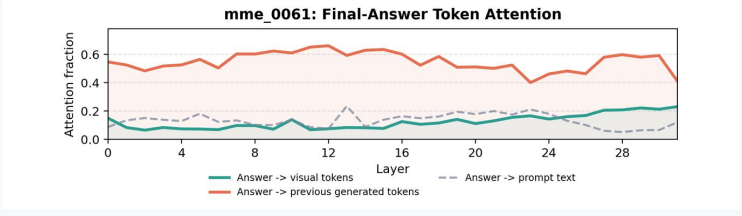
Read the heatmap as answer-slot behavior over diffusion steps: green denotes the correct answer token interval, red denotes the drifted answer-token interval, and blue marks positions decoded at each step.

Diagnostic error mechanism

The trace over-extends from generic art-style language to an unsupported attribution to Jean-Baptiste Morel, adding painter-biography content not grounded by the image. Answer drift: correct answer persists through step 68, then the final answer stabilizes as Yes from step 69.

Attention diagnostic: Final answer token `Yes` attends more to generated prefix than visual tokens on average (0.550 vs 0.125); late-layer prefix/visual ratio = 2.68x.





Interpretation is diagnostic: the pattern is consistent with self-anchoring on the generated prefix, not a standalone causal proof.

Figure 9: Case5