MACHINES SEE BETTER WHEN THEY IMAGINE: DIFFUSION IMAGINATION REVEALS BETTER VISUAL PRIORS FOR MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are typically evaluated with human-readable images, assuming that representations optimized for human perception are also optimal for machine reasoning. We challenge this assumption by studying whether intermediate images along a diffusion denoising trajectory, which acts like machine imagination, can serve as effective visual priors for visual question answering (VQA). Given an input image, we perform diffusion inversion and reconstruction under different conditional priors, and further derive residual difference maps that highlight question-induced visual changes. Evaluated across top tier MLLMs on our proposed benchmark *DV-Bench*, the integration of diffusion visual priors consistently improve MLLMs' performance, especially on reasoning-heavy tasks. Our analysis shows that the gains are not only driven by a universal denoising step or conditional priors, but also by task type, MLLM and diffusion backbone. The same diffusion prior can bring significant improvement for one MLLM while harming another, revealing a model-dependent "meat/poison" effect. These results suggest that diffusion priors as visual representation have great potential for enhance the visual reasoning of different MLLMs. The source code will be fully released upon paper acceptance.

1 INTRODUCTION

Machine visual intelligence depends on how visual information is represented and made available for reasoning (Bengio et al., 2013). Modern vision models encode images through learned feature vectors, spatial tokens, and multimodal embeddings (He et al., 2016; Dosovitskiy et al., 2021; Radford et al., 2021). These advances motivate a complementary question about the inputs to such models: can machine-generated representations of an image help a machine reason about its content? For multimodal large language models (MLLMs), an original image provides the visual evidence for answering a question, but additional representations may make relevant cues easier to use. We investigate a simple hypothesis: **machines can see better when their visual inputs are complemented by appropriate machine-generated features.**

We use *machine-generated visual features* to refer to image-form representations derived from an input image, including decoded intermediate diffusion states and residual maps. Their potential value lies in reorganizing visual evidence through a learned generative prior, making aspects of scene structure more accessible to a downstream MLLM. Diffusion models provide a controllable way to construct such representations (Labs, 2025; Stability AI, 2024). By inverting an image into a noisy latent state and reconstructing it under different textual conditions, we obtain a sequence of intermediate images along the denoising trajectory. These images may emphasize different aspects of the scene, including global layout, object relations, and question-relevant cues. Their usefulness, however, must be established empirically: generated features can be redundant or distort details needed to answer the question, and perceptual quality alone does not establish their value for reasoning.

The central challenge is therefore to determine whether these generated features offer useful visual representations, whether they complement the original image, and which configurations best support a given task and model. We organize this investigation around three research questions: *1) Standalone*

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.

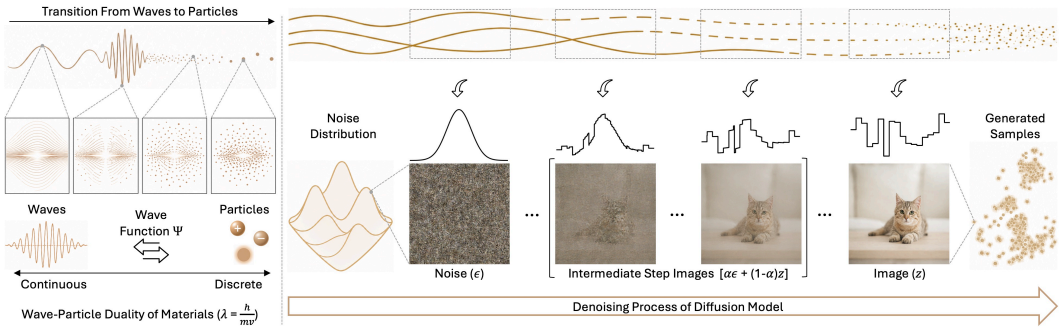

Figure 1: **From discrete visual samples to continuous diffusion imagination.** Human-readable images are only discrete samples from a richer underlying visual distribution. Left: We use the wave-particle analogy to illustrate how continuous representational processes can collapse into discrete observations. This analogy serves only as motivation and does not imply a physical or mathematical equivalence with diffusion models; Right: A diffusion model transforms a noise distribution into a final image through a sequence of intermediate denoising states. In our study, we obtain intermediate step images by inverting and reconstructing an input image under different textual conditions. We investigate whether these decoded intermediate images can provide more effective visual information for multimodal reasoning than the original image.

utility: How effective are machine-generated visual features as standalone inputs for MLLM reasoning? *2) Complementarity:* Can these features improve reasoning when combined with the original image? *3) Task–model dependence:* Which feature configurations are most useful for different tasks and MLLMs?

To address these questions, we study diffusion-generated features through visual question answering (VQA). We consider three representation types: reconstruction without text conditioning (R-N), reconstruction conditioned on the VQA question (R-Q), and a residual representation that captures visual differences between the two reconstructions (A-O). We define a *diffusion-prior configuration* (DPC) as a pairing of representation type and denoising step. Varying these configurations allows us to examine how textual conditioning and position along the denoising trajectory affect the usefulness of the generated features. We evaluate each configuration in two input settings: generated features alone and generated features combined with the original image. Comparing both settings against the original-image baseline distinguishes standalone utility from the gains obtained by supplementing the original visual input.

We evaluate multiple MLLMs on *DV-Bench*, a challenging benchmark covering scene understanding, spatial reasoning, counting, and object recognition. Our results show that appropriately chosen intermediate representations can improve VQA performance across different MLLMs. These gains depend on the task, representation type, and receiving model, with no single denoising step providing a universally optimal input. Different tasks benefit from different stages of the denoising trajectory, while different MLLMs respond differently to the generated representations. These findings motivate visual interfaces that supply machine-generated features according to their complementary value for the task and model at hand.

In this paper, our findings can be summarized as follows: 1) Appropriately selected intermediate images along diffusion trajectories provide effective visual priors for MLLM reasoning with broad generalizability. These benefits extend across MLLM families, diffusion backbones, and unfiltered benchmark subsets, with high response stability across repeated runs; 2) Different task types benefit from different levels of visual information, suggesting that diffusion trajectories expose a hierarchy of task-relevant evidence. Spatial reasoning generally favors earlier structural cues, whereas scene understanding benefits more from the richer semantic context available at later steps; 3) The usefulness of diffusion priors is distribution-dependent rather than universal. MLLMs exhibit distinct preferences for denoising steps and conditioning strategies, which we characterize as distributional resonance. These preferences also vary with the task and diffusion backbone, motivating model- and task-specific configuration selection; 4) The same intermediate representation can substantially benefit one MLLM while harming another, revealing a clear "meat/poison" phenomenon in machine perception. Such reversals can persist across denoising steps for a given conditional prior, suggesting systematic differences in how MLLMs interpret and use visual evidence.

2 Related Work

Human-inspired Visual Representation in MLLMs. Current MLLMs are largely developed (Singh et al., 2026; Gemini Team, Google, 2026; Anthropic, 2025; Liu et al., 2026; Bai et al., 2025; Team et al., 2026) and evaluated on human-readable visual data, where high-quality representations are implicitly associated with clear shapes, realistic colors, and captions grounded in human semantic concepts. While effective, this human-centric interface may impose a representational bottleneck for machine reasoning. Prior work on perceptual divergence shows that vision models can rely on cues that are weakly aligned with human perception, including texture bias (Geirhos et al., 2019), high-frequency signals (Wang et al., 2020), and adversarial features (Ilyas et al., 2019). Similar gaps also appear in transformer-based models such as ViT (Dosovitskiy et al., 2021) and CLIP (Radford et al., 2021), where stronger models do not necessarily become more aligned with human perceptual judgments (Hernández-Cámara et al., 2025) and can reason from visual structures that are difficult for humans to interpret (Naseer et al., 2021). Moreover, model attention and visual focus can be steered by language prompts (Gavrikov et al., 2025), suggesting that the optimal visual representation may depend on the reasoning context. These observations motivate our central question: whether a clear, human-readable image is always the best visual interface for MLLM reasoning.

Diffusion Trajectory as Visual Priors. Diffusion models have become a dominant paradigm for text-to-image generation (Rombach et al., 2022; Peebles & Xie, 2023; Podell et al., 2024; Esser et al., 2024; Black Forest Labs Team, 2024), evolving from DDPM/DDIM formulations (Ho et al., 2020; Song et al., 2021) toward flow-matching and rectified-flow models (Liu et al., 2023; Esser et al., 2024). Beyond synthesis, diffusion trajectories provide a hierarchical decomposition of visual data: earlier timesteps tend to capture coarse layout and global structure, while later steps recover finer semantic details (Liew et al., 2022; Yi et al., 2024; Tinaz et al., 2025). Intermediate diffusion representations have also been shown useful for downstream tasks such as zero-shot classification (Li et al., 2023a) and semantic correspondence (Tang et al., 2023; Luo et al., 2023). These findings raise a natural question: can intermediate diffusion representations, even less interpretable to humans, serve as more useful visual priors for machine visual reasoning?

3 Methods

3.1 Preliminaries

Diffusion Models. Diffusion models learn the transition from noise distribution P_{noise} to data distribution P_{data}. In this paper, we follow the flow matching (FM) framework. Let $z \in \mathbb{R}^d$ be the latent representation of an image I encoded via a pre-trained Variational Autoencoder (VAE), such that $z \sim P_{\text{data}}$. Let $\epsilon \sim \mathcal{N}(0, I_d)$ represent a sample from the standard Gaussian noise distribution and x_t denote the intermediate latent state at time $t \in [0, 1]$, we define the latent path as a linear interpolation $x_t = \alpha_t z + \sigma_t \epsilon,\ s.t.\ \alpha_0 = \sigma_1 = 0,\ \alpha_1 = \sigma_0 = 1$ where $\alpha_t, \sigma_t \geq 0$ are time-dependent schedulers that define the probability path. $v_\theta(x_t, t, c)$ denotes a time-dependent velocity predicted by the trained diffusion model with parameters θ, conditioned on text prompt c.

MLLMs. Let $\mathcal{M}_\phi$ denotes the MLLM parametrized by ϕ. Given an input image I and a natural language question Q, $\mathcal{M}_\phi$ aims to find the optimal answer sequence $A = \{a_1, a_2, \ldots, a_n\}$ that maximizes the conditional probability $A = \arg\max_{A'} P(A' \mid \mathcal{P}(\mathcal{E}(I)), \mathcal{T}(Q); \phi)$, where $\mathcal{P}, \mathcal{E}, \mathcal{T}$ are projector, vision encoder and tokenizer, respectively. For simplicity, we use $A = \mathcal{M}_\phi(I, Q)$ to represent this process in this paper.

3.2 Diffusion Roundtrip Process

The diffusion roundtrip process consists of two stages: deterministic inversion and denoising reconstruction. Starting from the latent representation of the input image I_{ori}, we first invert it towards a noisy intermediate state and then reconstruct it through the forward denoising dynamics.

Deterministic Inversion. Let $z_{\text{ori}} = \text{VAE}_{\mathcal{E}}(I_{\text{ori}})$ denote the latent state of the input image, where $\mathcal{E}$ refers to the encoder. Given an inversion strength parameter $s \in [0, 1]$, we define a starting timestep $\tau = 1 - s$. The inversion process obtains a noisy latent state x_τ by solving the ordinary differential equation (ODE) backward in time from $t = 1$ to τ:

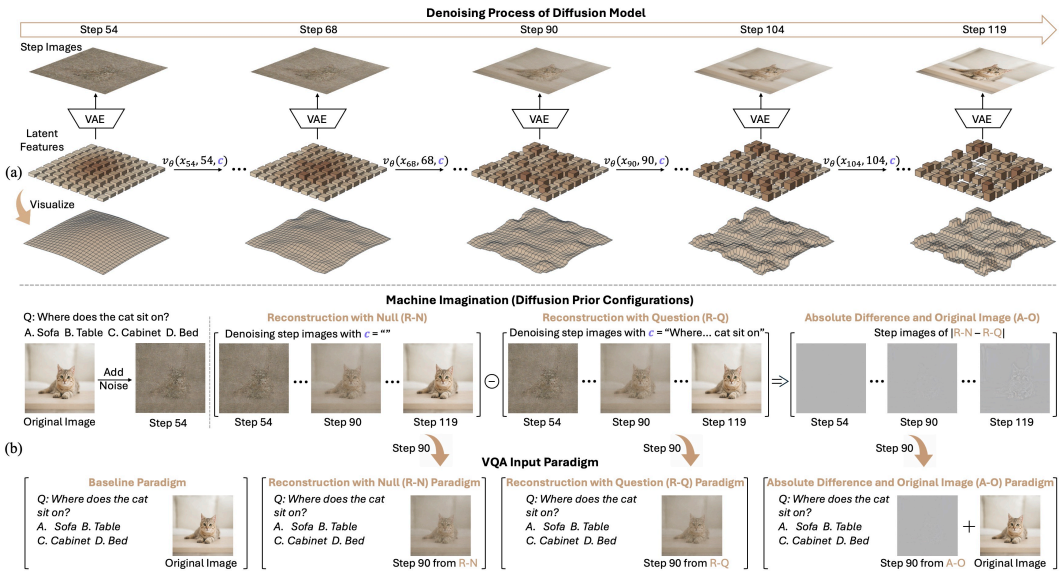

Figure 2: **Method overview.** Given an original image and a visual question, we first encode the image into the latent space and invert it to a noisy intermediate state. Starting from this state, we perform denoising with different conditioning priors: without text conditioning, which produces R-N; question as text conditioning, which produces R-Q; and the absolute difference between R-N and R-Q combined with the original image, which produces A-O. Intermediate images are decoded at multiple denoising steps, yielding diffusion-prior configurations (DPCs), where each DPC is defined by a conditional prior and a denoising step. These DPCs are then used as visual inputs to MLLMs for VQA, allowing us to evaluate whether diffusion imagination provides more effective visual evidence than the original image alone.

$$x_\tau = z_{\text{ori}} - \int_\tau^1 v_\theta(x_t, t, c)\, dt \xrightarrow{\text{Discretize}} z_{\text{ori}} - \sum_{i=1}^{K} \Delta t \cdot v_\theta(x_{t_i}, t_i, c) \tag{1}$$

where K is the number of discrete steps and Δt denotes the step size.

Denoising Reconstruction. The denoising phase attempts to recover the I_{ori} from the inverted latent state x_τ. This is achieved by solving the ODE forward in time, starting from x_τ at $t = \tau$ and integrating forward to $t = 1$:

$$\hat{z} = x_\tau + \int_\tau^1 v_\theta(x_t, t, c)\, dt \xrightarrow{\text{Discretize}} x_\tau + \sum_{j=1}^{K} \Delta t \cdot v_\theta(x_{t_j}, t_j, c) \tag{2}$$

where $\hat{z}$ denotes the reconstructed latent state and reconstructed image can be decoded using $I_{\text{rec}} = \text{VAE}_{\mathcal{D}}(\hat{z})$, where $\mathcal{D}$ refers to the decoder. During the denoising process, the evolution of the intermediate latent state x_t is computed iteratively starting from the inverted state x_τ For each discrete step, the subsequent state $x_{t+\Delta t}$ is expressed as $x_{t+\Delta t} = x_t + \Delta t \cdot v_\theta(x_t, t, c)$ with $x_{t=\tau} = x_\tau$. Intermediate reconstructions along the denoising trajectory can be obtained by decoding each latent state as $I_{\text{t}} = \text{VAE}_{\mathcal{D}}(x_t)$.

3.3 VQA WITH DIFFERENT DIFFUSION PRIOR CONFIGURATIONS

To investigate how DPCs influence the visual reasoning of MLLMs, we construct and analyze visual inputs along the denoising trajectory with three types of conditional priors: *reconstruction with null* (R-N), *reconstruction with question* (R-Q), and *absolute difference with original image* (A-O). Starting from the same inverted latent state x_τ obtained from the input image I_{ori}, we perform two denoising processes using different textual conditions. Specifically, let $c_\emptyset$ denote the null prompt and let c_Q denote the question Q. The corresponding latent trajectories are governed by

$$x^{\emptyset}_{t+\Delta t} = x^{\emptyset}_t + \Delta t\, v_\theta(x^{\emptyset}_t, t, c_\emptyset), \qquad x^{Q}_{t+\Delta t} = x^{Q}_t + \Delta t\, v_\theta(x^{Q}_t, t, c_Q), \tag{3}$$

with the shared initialization $x^{\emptyset}_\tau = x^{Q}_\tau = x_\tau$. At each timestep t, the corresponding decoded images of R-N and R-Q are given by $I^t_{\text{R-N}} = \text{VAE}_{\mathcal{D}}(x^{\emptyset}_t)$ and $I^t_{\text{R-Q}} = \text{VAE}_{\mathcal{D}}(x^{Q}_t)$, respectively. To further highlight the question-induced visual discrepancy, A-O integrates the original image with the pixel-wise

Table 1: **Results with validation-selected DPCs.** For each model and task category, we report the accuracy of the original-image baseline (Base), the average accuracy obtained using the selected top 1 and 3 DPCs on validation set, and its improvement over the baseline (Gain). From the result, diffusion priors consistently improve the performance across MLLMs and task types, with particularly strong gains on reasoning-heavy tasks.

		Spatial Reasoning			Scene Understanding			Counting			Object Recognition		
	Model	Base	DPC	Gain	Base	DPC	Gain	Base	DPC	Gain	Base	DPC	Gain
Top 1	GPT-5-Mini	57.26	68.77	+11.51	24.21	38.01	+13.80	29.52	42.17	+12.65	34.67	41.26	+6.59
	Claude-Haiku-4.5	33.42	36.16	+2.74	33.17	36.32	+3.15	28.31	34.34	+6.02	32.66	38.11	+5.44
	Gemini-3.1-Flash-Lite	43.56	49.86	+6.30	32.20	37.53	+5.33	33.73	39.76	+6.02	46.13	48.71	+2.58
	Kimi-K2.5	47.40	60.55	+13.15	35.84	43.58	+7.75	38.55	46.99	+8.43	50.43	54.44	+4.01
	Ministral-3-14B	32.33	40.00	+7.67	35.59	37.77	+2.18	28.31	33.73	+5.42	35.24	37.54	+2.29
	Qwen-3-VL-235B	42.19	46.58	+4.38	34.14	40.68	+6.54	33.73	37.95	+4.22	39.26	48.42	+9.17
Top 3	GPT-5-Mini	57.26	67.95	+10.68	24.21	37.21	+12.99	29.52	40.36	+10.84	34.67	40.69	+6.02
	Claude-Haiku-4.5	33.42	35.53	+2.10	33.17	35.35	+2.18	28.31	31.93	+3.61	32.66	36.58	+3.92
	Gemini-3.1-Flash-Lite	43.56	48.13	+4.57	32.20	35.35	+3.15	33.73	38.15	+4.42	46.13	47.76	+1.62
	Kimi-K2.5	47.40	59.00	+11.60	35.84	40.68	+4.84	38.55	44.38	+5.82	50.43	52.72	+2.29
	Ministral-3-14B	32.33	38.26	+5.94	35.59	36.32	+0.73	28.31	32.73	+4.42	35.24	36.87	+1.62
	Qwen-3-VL-235B	42.19	45.57	+3.38	34.14	39.47	+5.33	33.73	37.35	+3.61	39.26	47.66	+8.40

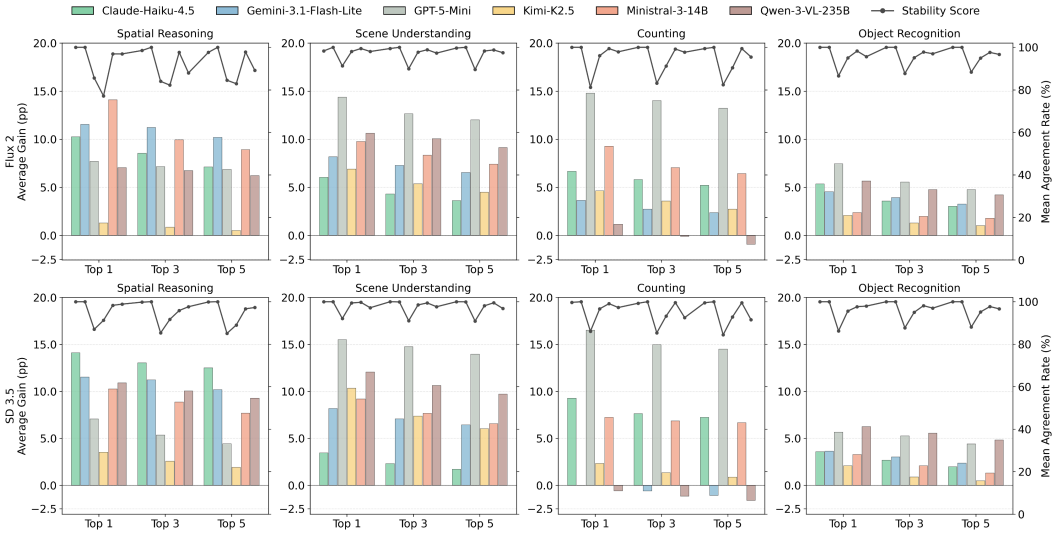

Figure 3: **Average performance gain and MAR of top 1, 3, 5 DPCs across MLLMs on *DV-Bench-tiny*.** Bar plots report the average accuracy gains, in percentage points, achieved by the top-1, top-3, and top-5 DPCs for each MLLM across four task types. Results are shown for both Flux 2 and SD 3.5. The line plots show the mean agreement rate across repeated runs, measuring response stability under the same DPC. The consistently positive gains and high stability scores indicate that intermediate diffusion priors provide robust improvements rather than isolated or stochastic benefits.

absolute difference between the two reconstructions $I^t_{\text{A-O}} = \left\{ \left| I^t_{\text{R-N}} - I^t_{\text{R-Q}} \right|, I_{\text{ori}} \right\}$. $\left| I^t_{\text{R-N}} - I^t_{\text{R-Q}} \right|$ denotes the pixel-level absolute difference map, which captures the visual changes induced by conditioning on Q, as shown in Fig. 2. Given input image $I \in \mathcal{I}^t$, s.t. $\mathcal{I}^t = \left\{ I^t_{\text{R-N}}, I^t_{\text{R-Q}}, I^t_{\text{A-O}} \right\}$, the VQA process of MLLM is formulated as $A = \mathcal{M}_\phi(I, Q)$. This formulation enables us to systematically examine how the MLLM's answer varies with intermediate reconstructions generated under different diffusion conditions, and in turn to assess whether question-conditioned diffusion priors expose task-relevant visual evidence that is not directly apparent in the original image.

4 Experiments

Experimental Setup. We evaluate whether intermediate step images along diffusion trajectory can serve as useful visual priors for MLLM reasoning. Our evaluation is conducted on *DV-Bench*, a hard VQA benchmark constructed based on A-OKVQA (Schwenk et al., 2022), OmniSpatial (Jia et al., 2025), SEED-Bench-H (Li et al., 2023b), VMCBench (Zhang et al., 2025), and VSTAR (Wu & Xie, 2023). To focus on cases where visual reasoning remains challenging, we first filter samples using GPT-4o-mini and Qwen-3-VL-32B and retain examples on which these models perform poorly. The

resulting benchmark contains 1.3K examples across four task categories: scene understanding, spatial reasoning, counting, and object recognition. For each VQA, we generate intermediate step images following the procedure described in Sec. 3. We use set the total denoising steps to be 120 and set the inversion strength to 0.55, which preserves substantial visual consistency between the reconstructed images and the original input while still exposing intermediate diffusion priors. Since this strength corresponds to starting the reconstruction from step D54, we save intermediate images approximately every 7 denoising steps from D54 to D119. We additionally construct a 40-sample validation set to select, for each MLLM and task type, the best performing DPC (*e.g.*let@tokeneonedotR-N+D54). For MLLMs, We evaluate top-tier MLLM families including GPT (Singh et al., 2026), Gemini (Gemini Team, Google, 2026), Claude (Anthropic, 2025), Kimi (Team et al., 2026), Qwen (Bai et al., 2025) and Ministral (Liu et al., 2026). Their versions are listed in Tab. 1. For diffusion models, we use Flux 2-Klein (Labs, 2025) (Flux 2) as the primary diffusion backbone and include Stable Diffusion 3.5-Medium (Stability AI, 2024) (SD 3.5) in Sec. 5 for analysis and findings.

Evaluation Protocol. For each MLLM and task type, we select the top 1 and 3 best-performing DPCs on the validation set and then report their average accuracy on *DV-Bench*. The baseline is the same MLLM evaluated using the original image of VQA.

Main Results. Tab. 1 shows that validation-selected DPC consistently improve MLLM performance across model families and task categories. The gains are especially clear on reasoning-heavy tasks. For spatial reasoning, GPT-5-Mini improves 11.51% and Kimi-K2.5 improves 13.15%. We also observe substantial improvements on perception-oriented tasks, such as Qwen-3-VL-235B's 9.17% improvement on object recognition. These results indicate that intermediate diffusion images are not merely visual perturbations: they can expose task-relevant evidence that is more accessible to certain MLLMs than the original human-readable images. More importantly, the improvements are not limited to a single model family. Across all tested MLLMs, almost every task category benefits from at least one DPC, which suggests that improving MLLM reasoning requires not only stronger models, but also visual representations better aligned with machine reasoning.

5 FINDINGS

The main results show that validation-selected DPCs can improve MLLM reasoning on *DV-Bench*. We now analyze where these improvements come from and why the optimal DPC varies across MLLMs and task types. To support fine-grained analysis, we construct *DV-Bench*-tiny, a 434-sample subset of *DV-Bench* with more than 100 examples for each task type. For each sample, we generate intermediate denoising images using both Flux 2 and SD 3.5, evaluate all DPCs across MLLMs, and repeat each setting three times. This enables analysis along three dimensions: denoising step, conditional priors, and diffusion backbone, along with stability. We quantify stability using the mean agreement rate (MAR), defined as the average fraction of repeated runs that agree with the majority answer for each query: $S_{\text{MAR}} = \frac{1}{N} \sum_{q \in Q} \frac{M_q}{R_q}$, where N is the number of VQAs, R_q is the number of repeated runs for VQA q ($R_q = 3$ in our setting), and M_q is the number of runs matching the majority answer. Higher values indicate more stable behavior under the same DPC.

5.1 GENERALIZABLE PERFORMANCE GAINS WITH HIGH STABILITY

In Fig. 3, we evaluate whether the improvements are caused by a single specific DPC or by a broader set of them. For each MLLM and task type, we report the average gain of the top-1, top-3, and top-5 DPCs, together with their MAR across repeated runs. The top-1 result reflects the best available DPC, while the top-3 and top-5 verifies whether the benefit persists beyond a single optimal configuration. The results show that the gains are not brittle. On reasoning-heavy task types such as scene understanding and spatial reasoning, several MLLMs achieve more than 10% improvement with the top-1 DPC. More importantly, many models still maintain clear gains when averaging over the top-3 or top-5 DPCs, indicating that useful diffusion priors occupy a nontrivial region of the DPC space rather than a single isolated point. This pattern appears under both Flux 2 and SD 3.5, suggesting that the effect is not tied to one specific diffusion backbone.

The MAR further support this conclusion. Across models, task types, and diffusion backbones, the improvements are highly reproducible across repeated runs, often exceeding 80% and in many

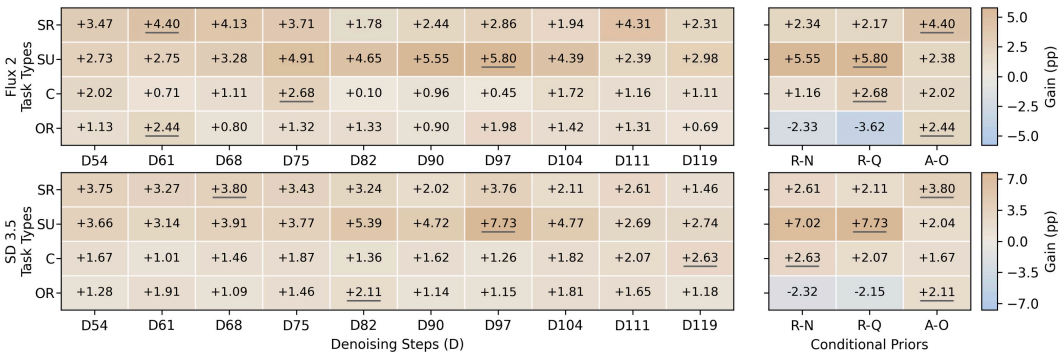

Figure 4: **Average improvements brought by different step images and methods for different task types across MLLMs.** SR, SU, C and OR refer to spatial reasoning, scene understanding, counting and object recognition, respectively. R-N, R-Q and A-O means reconstruct with null, reconstruct with question and absolute difference with original.

cases approaching or exceeding 90%. Thus, the observed gains are unlikely to be caused merely by stochastic answer variation. Instead, they reflect a stable improvement by diffusion priors.

5.2 Hierarchical and Conditional Diffusion Priors

Denoising Steps Expose Hierarchical Visual Priors. Along denoising trajectory, earlier steps tend to expose coarse global structure and scene layout, while later steps recover more detailed object appearance, semantic content, and visual coherence. Fig. 4 left illustrates how different denoising steps provide different types of visual evidence for MLLM reasoning. Spatial reasoning benefits most from relatively early steps (*e.g*let@tokeneonedotD54-75), suggesting that it relies more on layout, relative position, and coarse geometric relations than on fine texture. Scene understanding, in contrast, tends to peak at later steps (*e.g*let@tokeneonedotD82-104), where richer semantic context and coherent scene structure become more accessible. Counting and object recognition show weaker monotonic trends, likely because they depend on more fragile visual evidence: counting requires objects to remain separable, while object recognition often relies on local details that diffusion may smooth or alter. Overall, the denoising trajectory forms a hierarchy of machine-usable visual priors, with different steps exposing task-specific evidence.

Conditional Priors Shape Visual Evidence. Fig. 4 right compares how different conditional priors shape visual evidence to the MLLM. R-N and R-Q are reconstruction-based priors: R-N exposes the diffusion model's default reconstruction of the input, while R-Q uses the question to guide which semantic structures are emphasized during denoising. These priors are especially useful for scene understanding, where re-rendering can make ambiguous visual evidence more canonical and semantically coherent. However, they can be less reliable for tasks requiring fine-grained perception such as object recognition, since reconstruction may smooth small objects, alter subtle textures, or introduce plausible but incorrect details. A-O provides a complementary residual prior: it preserves the original image while using the difference between null- and question-conditioned reconstructions as a question-induced saliency signal. This makes A-O useful when both grounding and focus are needed, such as in spatial reasoning and object recognition. Together, these results suggest that reconstruction-based priors help when semantic reorganization is beneficial, while residual priors help when original-image fidelity and question-guided saliency are both required.

5.3 Distributional Resonance Between MLLMs and DPCs

As illustrated in Fig. 5, the usefulness of a DPC is not an intrinsic property of the intermediate image alone. Instead, it depends on the alignment among the task type, denoising step, conditional prior, diffusion backbone, and the MLLM. We refer to this alignment as distributional resonance.

Resonance Between MLLMs and Conditional Priors. We first examine the conditional-prior results in Fig. 5. For each task, the top performing conditional priors are not randomly distributed across MLLMs, but instead form several shared preference groups. In spatial reasoning, some MLLMs consistently benefit more from reconstruction-oriented priors, while others obtain larger

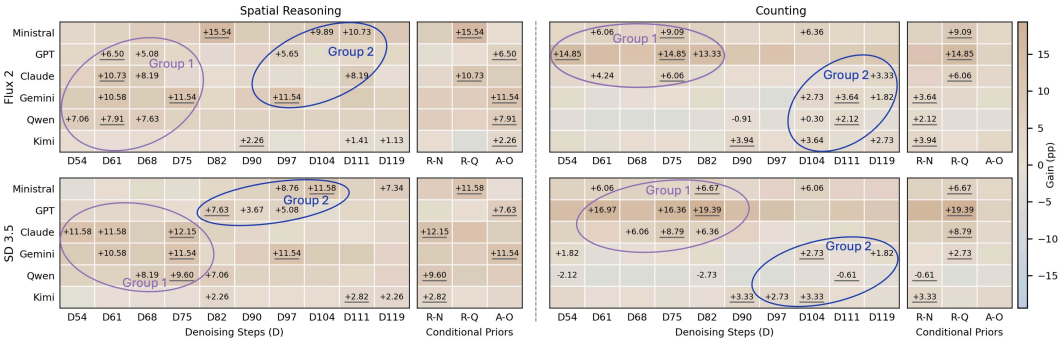

Figure 5: **Performance gains across denoise steps and conditional priors across MLLMs on spatial reasoning and counting.** The detailed versions of used MLLMs are illustrated in Tab. 1. For each task and diffusion backbone, we separately report gains across denoising stages and conditional priors. To emphasize, we only show the value of top 3 denoise steps and top 1 conditional priors with color and underline the highest gain for each MLLM. The highlighted groups reveal shared resonance patterns across MLLMs: models within the same group tend to favor similar denoising regions, while the top-1 conditional priors also concentrate into a small number of shared preference groups. The results reveal that the useful DPC depend jointly on the MLLM, task type, diffusion backbone, denoising stage, and conditional prior, indicating that diffusion-prior effectiveness is distribution-dependent rather than universal. We annotate Full version in Appendix.

gains from A-O. This suggests that spatial reasoning requires not only useful visual evidence, but also an evidence format that matches the model's ability to decode spatial layouts and object relations. A similar pattern appears in counting: although counting depends on fine-grained local evidence, the best conditional priors still concentrate around a small subset of prior types, indicating shared evidence-formatting preferences among different MLLMs. Therefore, conditional-prior resonance characterizes *how* diffusion-derived evidence should be formatted for different groups of MLLMs.

Resonance Between MLLMs and Denoising Steps. We next study the denoising-step results. The highlighted Group 1 and Group 2 in Fig. 5 represent groups of MLLMs that share similar preferred denoising regions. This shows that the optimal denoising step is not universal, but depends on the interaction between the task, diffusion backbone, and MLLM. For spatial reasoning, group 1 tends to gain more from earlier or intermediate steps that preserve structural layout information, while group 2 benefits more from later steps with more visually resolved details. For counting, similar groups can also be observed, but the distinction is less sharp because counting relies on local object evidence that may be clarified or distorted throughout denoising. The grouping pattern also has slight changes across Flux 2 and SD 3.5, suggesting that different diffusion backbones expose different visual evidence at the same denoising step. Overall, denoising-step resonance characterizes *where* along the denoising trajectory different groups of MLLMs can best extract useful task-relevant evidence.

5.4 Reversed Distribution: One Model's Meat is Another Model's Poison

Fig. 6 examine the most extreme cases by comparing the top 1 and 3 highest- and lowest-performing DPCs for each MLLM. While Fig. 5 shows that different MLLMs prefer different regions of the DPC space, Fig. 6 reveals a stronger phenomenon: the same DPC can act as the most useful visual evidence for one MLLM while becoming the most harmful distractor for another. We refer to this model-dependent reversal as the "meat/poison" phenomenon.

Reversal Among Denoising Steps. As shown in Fig. 6(a), the same denoising step can appear among the best DPCs for some MLLMs (yellow box) and among the worst for others (purple box). This suggests that denoising steps are not globally informative or noisy. Instead, each step exposes a particular mixture of coarse layout, residual noise, semantic structure, and recovered detail. For one MLLM, this mixture may align with the distribution expected by its visual encoder and projector, making relevant evidence easier to decode. For another, the same mixture may suppress needed cues or introduce distracting patterns. Thus, an intermediate step image can be information-rich but evidence-poor for a particular model. Earlier steps may help MLLMs that reason from coarse layout and object relations, but harm others that rely on resolved boundaries and local texture.

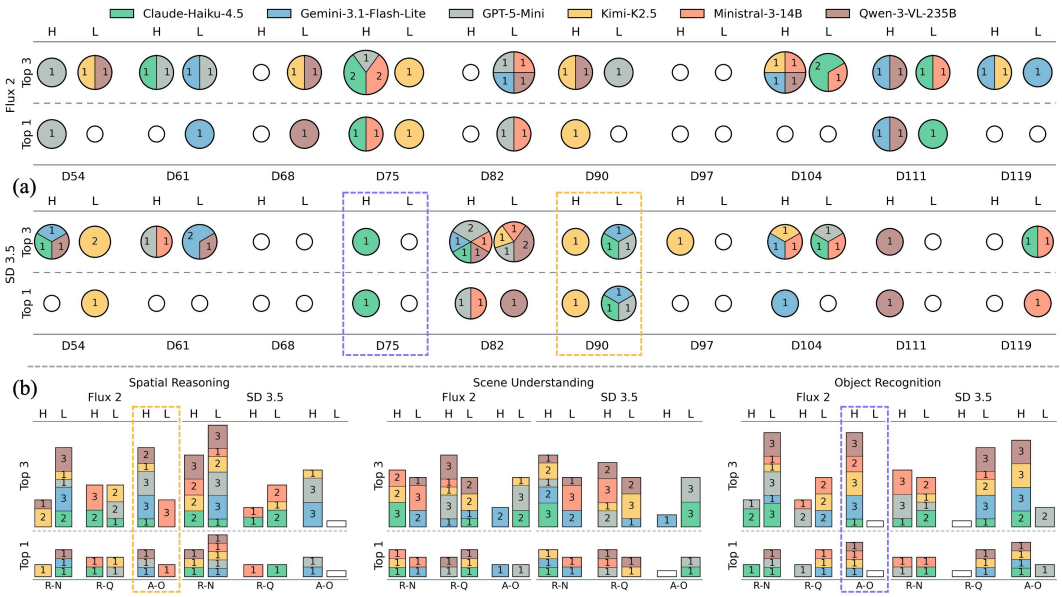

Figure 6: **Top 1 and 3 highest (H) and lowest (L) performing DPCs across MLLMs.** Numbers inside circles or bars indicate how often a model selects the corresponding denoising step or conditional prior among its highest- or lowest-performing DPCs. (a) shows step-level statistics for counting, where the same denoising step may be helpful for one model but harmful for another. (b) shows the conditional prior-level statistics for other three task types. We highlight some steps and methods with or without model reversal using boxes.

Reversal Among Conditional Priors. Fig. 6(b) shows that this reversal also appears across conditional diffusion priors. Compared with step-level behavior, prior-level preferences are often more concentrated: for a given model and task, the highest- or lowest-performing DPCs tend to cluster around particular prior types. This indicates that the conditional prior itself can act as a stable interface bias: once a model is compatible or incompatible with a certain evidence format, this tendency can persist across multiple denoising steps. This pattern reveals a stronger form of meat/poison behavior: the same prior can be consistently beneficial or consistently harmful across steps. This suggests that the dominant factor is not the amount of information contained in the prior, but the compatibility between the prior's evidence format and the MLLM's visual interface. When this format matches the MLLM's perceptual bias, it becomes a robust guide, but when it conflicts with that bias, it becomes a persistent distractor. Thus, DPC effectiveness is determined not only by the information contained in the diffusion prior, but also by its compatibility with the receiving MLLM.

6 Conclusion and Future Work

In this work, we demonstrate that machine imagination, instantiated as intermediate step images along the denoising trajectory of a diffusion model, can serve as effective visual priors for enhancing the visual reasoning of MLLMs. Experiments across diverse MLLMs on our proposed dataset *DV-Bench* indicates consistently improve performance, especially on reasoning-heavy tasks. Our analysis further shows that their usefulness is not universal: different task types, diffusion backbones, and MLLMs prefer different diffusion visual priors and the same prior can help one model but harm another, revealing a clear meat/poison effect. These findings suggest that better visual reasoning may require not only stronger models, but also visual representations better aligned with machine perception. While our results demonstrate the promise of diffusion visual priors for VQA, it may impair fine-grained recognition or detection, particularly for tiny object detection as denoising preserves global, low-frequency structure while suppressing or altering local, high-frequency details. Occasionally, it may also alter, omit, or replace objects, resulting in denoising hallucinations. Besides, their generalization to action planning based on vision language action framework remains unclear, which we will explore in the future work.

AI USE STATEMENT

In this work, we used generative AI tools for assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools for generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, clean and reformat dataset and formulate mathematical claims, provide critical ingredients for proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We thoroughly checked all LLM-generated texts inside this paper and LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper has no ethical concerns.

REPRODUCIBILITY STATEMENT

The code will be open sourced upon acceptance.

REFERENCES

Anthropic. Introducing claude haiku 4.5. Anthropic News, 2025. URL https://www.anthropic.com/news/claude-haiku-4-5.

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

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL https://doi.org/10.1109/TPAMI.2013.50.

Black Forest Labs Team. Announcing flux.1. Black Forest Labs Blog, 2024. URL https://blackforestlabs.ai/announcing-black-forest-labs/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Stephanie Fu, Netanel Y. Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: learning new dimensions of human visual similarity using synthetic data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc., 2023.

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Can we talk models into seeing the world differently? In *The*

Thirteenth International Conference on Learning Representations (ICLR), 2025. URL https://arxiv.org/abs/2403.09193.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Gemini Team, Google. Gemini 3.1 pro - model card. Google DeepMind Technical Report, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

Pablo Hernández-Cámara et al. Do vision transformers see like humans? evaluating their perceptual alignment. *arXiv preprint arXiv:2508.09850*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models, 2025. URL https://arxiv.org/abs/2506.03135.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217, October 2023a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023b.

Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière,

Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gWxpdtQpiYV.

Ziyi Luo, Ziang Luan, et al. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, et al. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 146–162, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20074-8.

Aaditya Singh et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2026. URL https://arxiv.org/abs/2601.03267.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Stability AI. Introducing stable diffusion 3.5, 2024. URL https://stability.ai/news/introducing-stable-diffusion-3-5.

Lunjun Tang, Ning Jia, Jiaya Jia, et al. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao,

Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. *arXiv preprint arXiv:2504.15473*, 2025.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8684–8694, 2020.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL https://arxiv.org/abs/2312.14135.

Xuanyu Yi, Zike Wu, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, and Hanwang Zhang. Diffusion time-step curriculum for one image to 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9948–9958, June 2024.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation, 2025. URL https://arxiv.org/abs/2501.03225.

Machines See Better When They Imagine: Diffusion Imagination Reveals Better Visual Priors for Multimodal Reasoning

Supplementary Material

CONTENTS

A Full Visualizations **15**

B Additional Experiments **17**
B.1 Evaluation on Unfiltered Benchmarks 17
B.2 Reconstruction Fidelity and Task Performance 17
B.3 Additional Baselines 18
B.4 Computational Cost 18

C VQA Inputs **20**
C.1 Prompt Template 20
C.2 VQA Examples 21

A FULL VISUALIZATIONS

Fig. 7 and 8 extend the model-level analysis in Sec. 5 to all four task categories. These visualizations use the 434 examples in *DV-Bench*-tiny and include all six MLLMs under both Flux 2 and SD 3.5. We follow the reconstruction settings in Sec. 4, with a 120-step diffusion schedule, an inversion strength of 0.55, and intermediate images sampled from D54 to D119. Performance changes are measured relative to the corresponding original-image baseline in percentage points; positive and negative values indicate improvements and degradations, respectively.

Dependence on the denoising step. Fig. 7 shows that the preferred denoising region varies across models, even within the same task category. For example, in spatial reasoning with Flux 2, Claude-Haiku-4.5 attains its largest displayed gain at D61, whereas Ministral-3-14B peaks at D82. The preferred region can also shift with the diffusion backbone: the peak for Ministral-3-14B moves to D104 under SD 3.5. Counting further illustrates the variation in model responses, with broadly positive gains for GPT-5-Mini but predominantly negative gains for Qwen-3-VL-235B in the displayed step profiles. Thus, task-level trends summarize heterogeneous model responses and should not be interpreted as a universal prescription for selecting a denoising step.

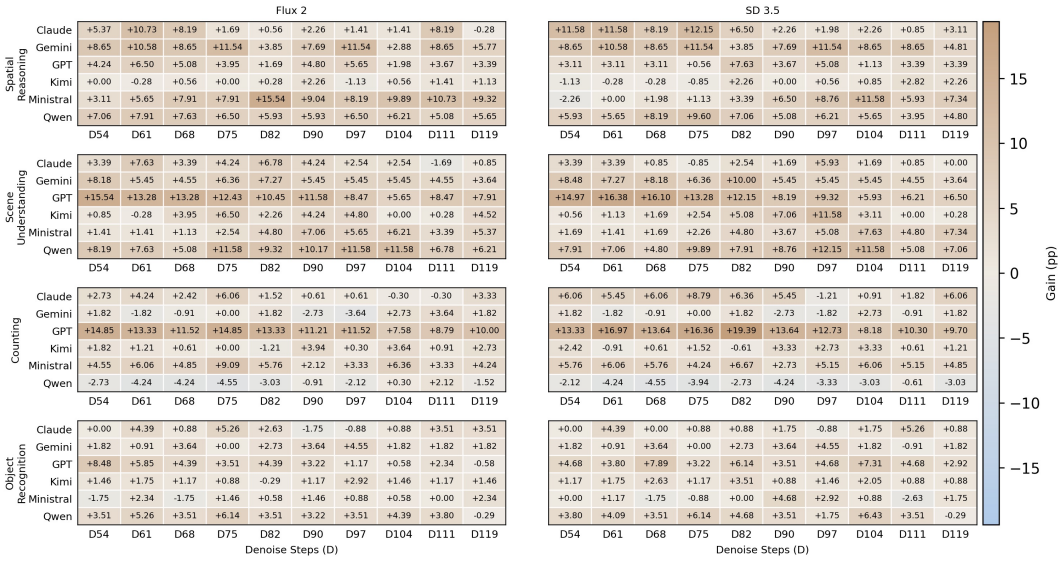

Figure 7: **Performance gains across denoising steps for different MLLMs, tasks, and diffusion backbones.** We report the gain of each MLLM under different denoising steps for four task categories: spatial reasoning, scene understanding, counting, and object recognition. Results are shown separately for Flux 2 and SD 3.5, with rows denoting MLLMs and columns denoting denoising steps. The color intensity indicates the magnitude of performance gain in percentage points. The results show that the most effective denoising step is not universal across MLLMs, tasks, or diffusion backbones. Instead, MLLMs exhibit step-level resonance patterns, where different groups of models obtain higher gains from different regions of the denoising trajectory.

Dependence on the conditional prior. Fig. 8 provides the complementary comparison across R-N, R-Q, and A-O. Some preferences persist across diffusion backbones: GPT-5-Mini favors R-Q for scene understanding and counting, while Gemini-3.1-Flash-Lite favors A-O for spatial reasoning and object recognition. The object-recognition results also illustrate the value of retaining the original image. For Kimi-K2.5 with Flux 2, the displayed gains are -2.63 and -6.43 percentage points for R-N and R-Q, respectively, compared with $+2.92$ for A-O. However, A-O is not uniformly preferable: GPT-5-Mini benefits more from a reconstruction-based prior on object recognition under both backbones. These observations support selecting the visual input format jointly with the receiving MLLM and task, rather than assuming that either reconstruction or residual evidence is consistently superior.

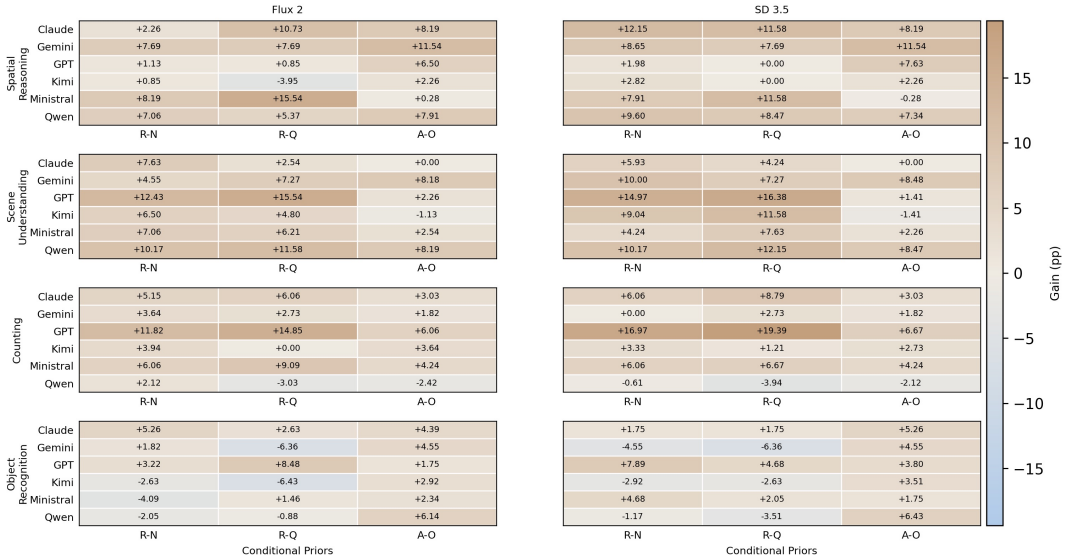

Figure 8: **Performance gains across conditional priors for different MLLMs, tasks, and diffusion backbones.** We report the gain of each MLLM under different conditional priors for spatial reasoning, scene understanding, counting, and object recognition. Results are shown for Flux 2 and SD 3.5, with rows denoting MLLMs and columns denoting conditional priors. The color intensity represents the gain in percentage points. Compared with denoising steps, conditional priors reveal evidence-formatting preferences across MLLMs: the best-performing priors are not randomly distributed, but often concentrate into a small number of shared preference groups. This suggests that different MLLMs may share similar prior-level resonance patterns when decoding diffusion-derived visual evidence.

B Additional Experiments

B.1 Evaluation on Unfiltered Benchmarks

Experimental protocol. To examine whether the benefits extend beyond the difficult examples retained in *DV-Bench*, we evaluate all six MLLMs on 600 randomly sampled examples from each of OMNI-Spatial and SEED-Bench-H. For both datasets, we do not separate the data samples into four task types and use different DPCs for each task type. Instead, we use one DPC, selected based on a validation set with 20 samples of each dataset for each model, and apply it through all 600 evaluation samples. In both experiments, the baseline uses the same MLLM with the original image, and DPC selection is performed at the model level rather than separately for each evaluation question.

Results and scope. Tab. 2 shows positive gains for all six MLLMs on both evaluation subsets. The mean improvement across models is 5.74 percentage points on OMNI-Spatial and 4.64 percentage points on SEED-Bench-H. On OMNI-Spatial, gains range from 3.65 to 7.66 percentage points, with Kimi-K2.5 improving from 48.17% to 55.83%. On SEED-Bench-H, gains range from 2.33 to 7.16 percentage points, with Claude-Haiku-4.5 improving from 51.67% to 58.83%. These results indicate that the observed benefits are not confined to the failure-filtered examples in *DV-Bench*. They establish improvements on the randomly sampled benchmarks.

Table 2: **Results with validation-selected DPCs on OMNI-Spatial and SEED-Bench-H.** For each model and benchmark, we report the original-image baseline accuracy (Base), the accuracy obtained using the DPC selected on the validation set (DPC), and the improvement over the baseline in percentage points (Gain). All accuracies are reported as percentages.

	OMNI-Spatial (600)			SEED-Bench-H (600)		
Model	Base	DPC	Gain	Base	DPC	Gain
GPT-5-Mini	47.50	54.67	+7.17	67.82	71.83	+4.01
Claude-Haiku-4.5	36.67	42.77	+6.10	51.67	58.83	+7.16
Gemini-3.1-Flash-Lite	48.17	52.35	+4.18	70.83	74.00	+3.17
Kimi-K2.5	48.17	55.83	+7.66	72.17	74.50	+2.33
Ministral-3-14B	34.00	39.67	+5.67	63.16	68.34	+5.18
Qwen-3-VL-235B	41.00	44.65	+3.65	63.67	69.67	+6.00

B.2 Reconstruction Fidelity and Task Performance

Settings and measurement. We assess reconstruction fidelity on data samples in *DV-Bench*-tiny using both Flux 2 and SD 3.5. The inversion strength is 0.55, reconstruction starts at D54 of the 120-step schedule, and the text guidance scale is set to 1. These settings are intended to limit departures from the input image while allowing the two conditioning branches to expose different visual evidence. R-N uses a null text prompt, and R-Q uses the question without answer options. Tab. 3 reports the mean DreamSim similarity between the final reconstruction and the original image for each task category. The parenthesized values reproduce the condition-level accuracy changes averaged across six MLLMs in Fig. 4. They summarize the DPC analysis and should not be interpreted as accuracy gains measured exclusively on the final reconstructions.

Table 3: **Image similarity and performance changes on *DV-Bench*-tiny.** Each entry reports the mean DreamSim similarity (Fu et al., 2023) between final reconstructions and their original images. Parentheses show the corresponding condition-level accuracy change from Fig. 4, averaged across six MLLMs and expressed in percentage points. Similarity is measured at the final reconstruction, whereas the performance summary concerns the sampled DPCs.

Model	Condition	Spatial Reasoning	Object Recognition	Scene Understanding	Counting
Flux 2	R-N	0.968 (+2.34)	0.960 (-2.33)	0.950 (+5.55)	0.965 (+1.16)
Flux 2	R-Q	0.940 (+2.17)	0.932 (-3.62)	0.925 (+5.80)	0.938 (+2.68)
SD 3.5	R-N	0.938 (+2.61)	0.921 (-2.32)	0.927 (+7.02)	0.938 (+2.63)
SD 3.5	R-Q	0.937 (+2.11)	0.918 (-2.15)	0.926 (+7.73)	0.930 (+2.07)

Task-dependent benefits and failure modes. The reported similarities range from 0.9178 to 0.9675, but the associated performance trends differ substantially by task. Scene understanding improves under all four backbone–condition combinations, with gains of 5.55–7.73 percentage points, whereas object recognition declines by 2.15–3.62 percentage points. For example, Flux 2 with R-N yields a final-image similarity of 0.9597 for object recognition alongside a condition-level accuracy change of -2.33 percentage points. This contrast shows that high aggregate perceptual similarity can coexist with degraded task performance. A plausible explanation is that reconstruction preserves broad scene structure while altering small objects, local textures, or other details needed for fine-grained recognition. By retaining the original image, A-O preserves access to these cues; its positive average object-recognition gains in Fig. 4 are consistent with this interpretation.

DreamSim provides an aggregate measure of perceptual fidelity, but these measurements do not quantify object insertion, removal, or attribute changes, nor do they establish fidelity at every intermediate step. Consequently, the similarity results alone cannot rule out reconstruction hallucinations or establish the mechanism behind the accuracy gains. The observed task dependence motivates treating diffusion reconstructions as alternative visual evidence whose utility requires validation for the intended task.

B.3 Additional Baselines

To examine whether simple image perturbations yield improvements comparable to our method, we evaluate two additional baselines on *DV-Bench*-tiny, which contains 434 samples, using six MLLMs. The first baseline adds random Gaussian noise to the original images, while the second randomly sharpens or blurs their edges. These perturbations modify image appearance without using a pretrained diffusion model. Tab. 4 compares their VQA accuracy with that obtained using the original images and the Top 1 and Top 3 results of our method. Values in parentheses indicate accuracy changes relative to the original-image baseline, expressed in percentage points.

The results show that neither perturbation baseline provides consistent improvements across models: each improves accuracy for only three of the six MLLMs. Random Gaussian noise produces changes ranging from -9.77 to $+4.54$ percentage points, while random edge sharpening or blurring yields changes between -1.65 and $+2.07$ percentage points. In contrast, our method improves accuracy across all six MLLMs, with gains of 3.72–11.07 percentage points for Top 1 and 2.77–9.85 percentage points for Top 3. Both settings also outperform both perturbation baselines for every model. These findings demonstrate the superiority of our method over the evaluated simple image perturbations, yielding larger and more consistent performance improvements across all six MLLMs.

Table 4: **Comparison with image perturbation baselines on *DV-Bench*-tiny.** Accuracy (%) is reported for original images, images perturbed with random Gaussian noise, and images with randomly sharpened or blurred edges. Top 1 and Top 3 report the corresponding results of our method, as illustrated in Tab. 1. Parentheses indicate accuracy changes relative to the original images, expressed in percentage points.

Model	Original	Random Noise	Random Edge	Top 1	Top 3
GPT-5-Mini	23.34	27.88 (+4.54)	22.35 (-0.99)	34.41 (+11.07)	33.19 (+9.85)
Gemini-3.1-Flash-Lite	35.14	34.33 (-0.81)	35.48 (+0.34)	42.12 (+6.98)	41.43 (+6.29)
Qwen-3-VL-235B	34.29	31.80 (-2.49)	34.33 (+0.04)	40.41 (+6.12)	39.65 (+5.36)
Kimi-K2.5	40.42	30.65 (-9.77)	39.86 (-0.56)	44.14 (+3.72)	43.19 (+2.77)
Ministral-3-14B	31.80	32.33 (+0.53)	33.87 (+2.07)	40.68 (+8.88)	38.63 (+6.83)
Claude-Haiku-4.5	30.91	31.57 (+0.66)	29.26 (-1.65)	37.99 (+7.08)	36.47 (+5.56)

B.4 Computational Cost

Tab. 5 reports diffusion preprocessing times measured on two NVIDIA RTX 3090 GPUs. The reported pipeline includes inversion, both reconstruction branches, residual-map computation, and VAE encoding and decoding. Its average total processing time is 43.43 seconds per sample for Flux 2 and 21.20 seconds for SD 3.5, making SD 3.5 approximately $2.05\times$ faster under the measured setup. Inversion and the two denoising branches account for most of the listed runtime: together they require 39.57 seconds for Flux 2 and 17.30 seconds for SD 3.5. Residual-map computation takes

approximately 3.1 seconds for either backbone, while the listed VAE operations are comparatively inexpensive. These measurements characterize image preprocessing; they do not provide the end-to-end latency of MLLM inference or the cost of validation-based DPC selection.

Table 5: **Computational cost of diffusion-based image processing.** We report the average total processing time per sample and the runtime of each processing stage for Flux 2 and SD3.5, measured using two NVIDIA RTX 3090 GPUs. R-N denotes denoising without a text prompt, while R-Q denotes denoising conditioned on the question. All times are in seconds.

Stage	Flux 2 (s)	SD3.5 (s)
Average total time per sample	43.43	21.20
Inversion (adding noise)	13.04	5.71
R-N (denoising without a text prompt)	13.23	5.79
R-Q (denoising conditioned on the question)	13.30	5.80
Absolute difference computation between R-N and R-Q	3.08	3.11
VAE encoding	0.08	0.08
VAE reconstruction decoding and saving	0.13	0.18

The preprocessing overhead is substantial relative to directly supplying the original image. A fixed R-N or R-Q configuration requires only its corresponding denoising branch, whereas A-O requires both branches to construct the residual map. Likewise, a fixed DPC requires only the selected visual input to be passed to the MLLM, although the preceding diffusion steps are still needed to obtain that input. These distinctions identify opportunities to reduce unnecessary computation after validation-based selection, but the table reports the measured pipeline and does not quantify such savings. The current results therefore demonstrate an accuracy benefit with additional computational cost; applications with strict latency requirements would require further efficiency evaluation.

C VQA INPUTS

C.1 PROMPT TEMPLATE

Prompts C.1 and C.2 specify the visual inputs and response format used for multiple-choice VQA. R-N and R-Q each supply one decoded reconstruction, obtained using a null prompt or the question as the diffusion condition, respectively. A-O supplies two images: the original image and the pixel-wise absolute difference between the R-N and R-Q reconstructions at the same denoising step. All visual inputs are presented in pixel space. The question and answer options are supplied to the MLLM, whereas only the question, without the options, is used to condition R-Q. The response schema requests an option index, its exact text, and a brief rationale.

Prompt C.1: R-N and R-Q Prompt Template

```
You are solving a multiple-choice VQA question.
Return ONLY valid JSON with this schema:
{"answer_index": <int>, "answer_text": <string>, "reasoning":
<short string>}

Rules:
- One image is provided: a denoising-stage reconstruction image
from the {mode} branch.
- This image corresponds to step label {step_label} (step index
{step_index}), file {step_image_name}.
- answer_index must be an integer between 0 and {max_index}.
- answer_text must exactly match one option string.

Question: {question}
Options: {options}
```

Prompt C.2: A-O Prompt Template

```
You are solving a multiple-choice VQA question.
Return ONLY valid JSON with this schema:
{"answer_index": <int>, "answer_text": <string>, "reasoning":
<short string>}

Rules:
- mode: {mode}
- Two images are provided:
1) original scene image: {original_image_name}
2) pixel-wise absolute difference between the R-N and R-Q
reconstructions at the same step: {step_image_name}
- The difference image corresponds to step label {step_label} (step
index {step_index}).
- answer_index must be an integer between 0 and {max_index}.
- answer_text must exactly match one option string.

Question: {question}
Options: {options}
```

C.2 VQA EXAMPLES

Fig. 9–12 present selected examples in which the use of DPCs corrects an erroneous original-image prediction (wrong to correct) across different task types. Each panel reports the question, answer options, ground-truth (GT) answer, conditional prior, denoising step, and predictions for the original and transformed inputs. Each setting is evaluated three times, as indicated by the repetition annotations such as $3/3$.

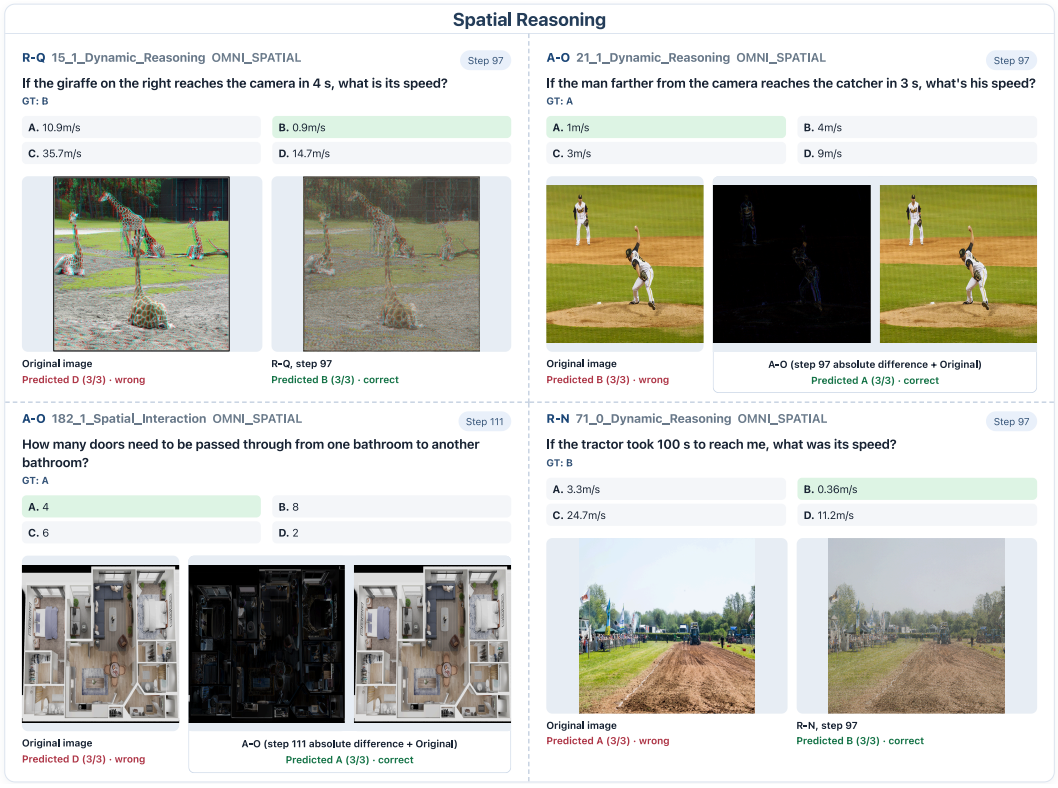

Figure 9: **Spatial reasoning examples.**

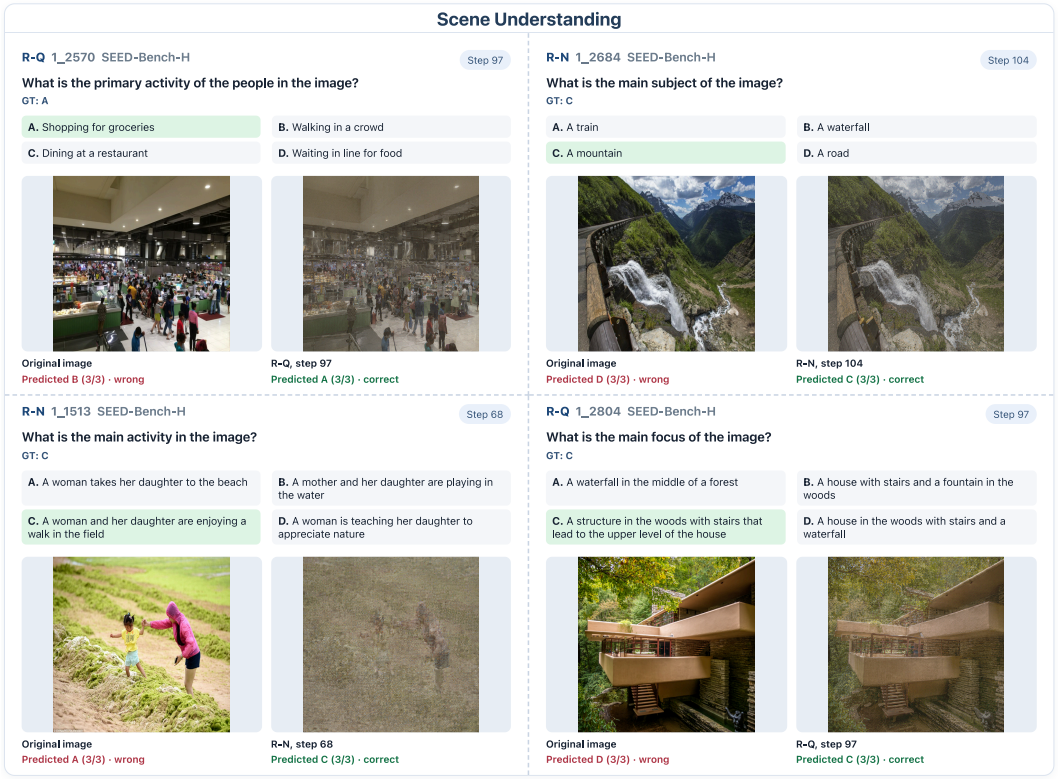

Figure 10: **Scene understanding examples.**

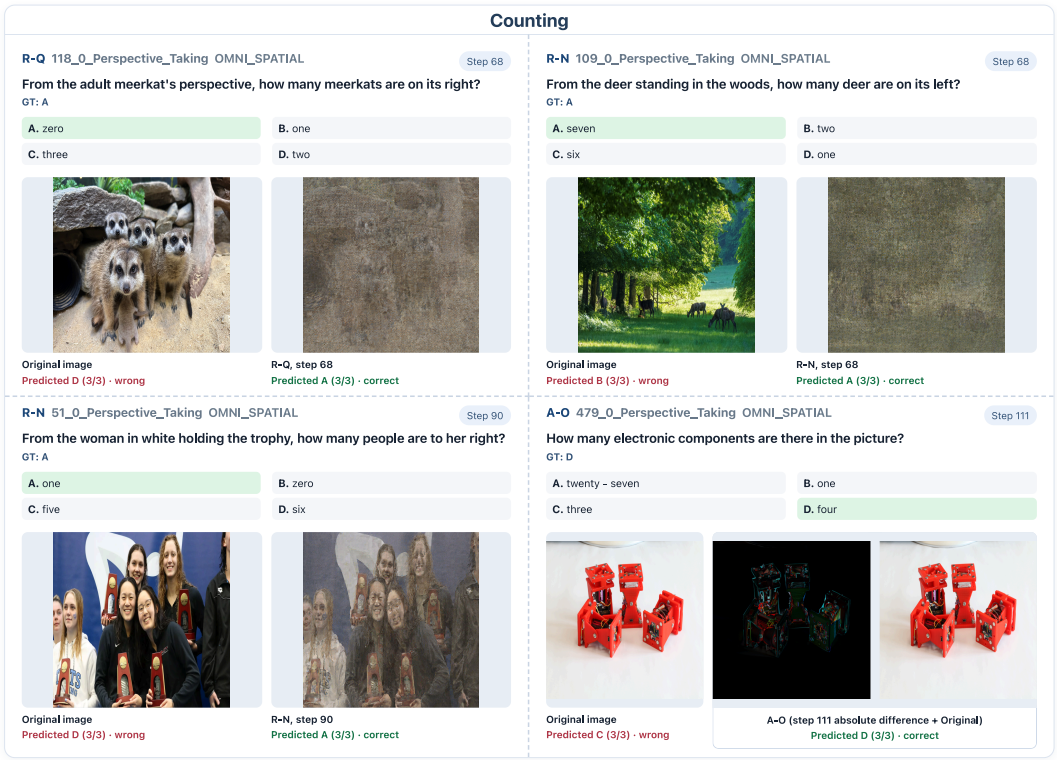

Figure 11: **Counting examples.**

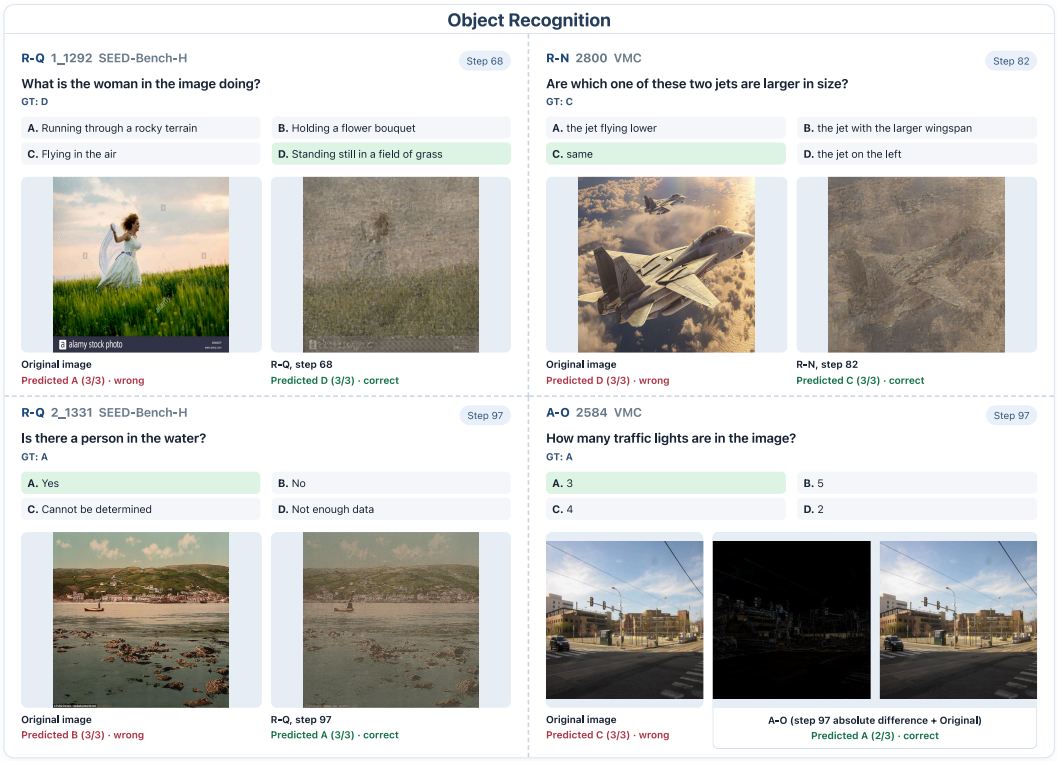

Figure 12: **Object recognition examples.**