CRYSTAL: SEEING LESS TO LEARN BETTER VISUAL LATENTS

Anonymous authors
Paper under double-blind review

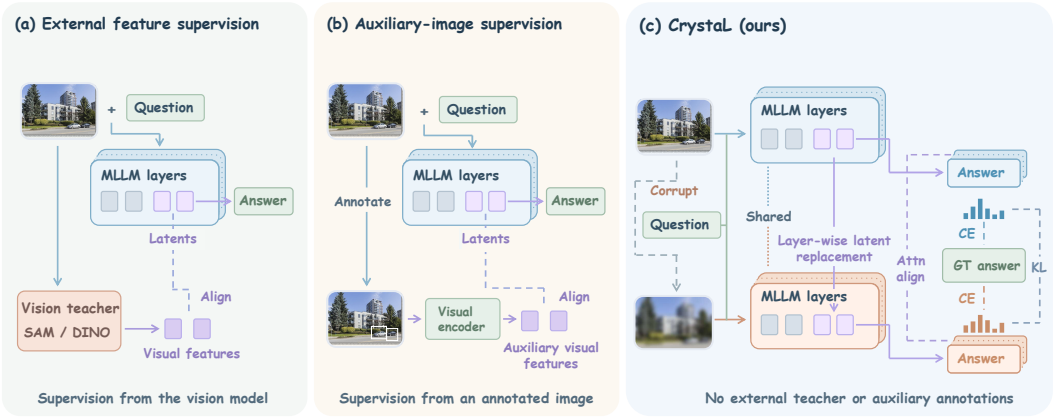


Figure 1: Supervision sources for latent visual reasoning. (a) The same input image is processed by external vision teachers, such as SAM(Kirillov et al., 2023) and DINO(Caron et al., 2021), to provide feature targets for latent alignment, as in CoVT(Qin et al., 2025). (b) An auxiliary image derived from annotations on the original image is encoded into visual targets for latent alignment, illustrating intermediate visual supervision such as Monet(Wang et al., 2025b). (c) CrystaL transfers latent states to a corrupted-image path through layer-wise replacement. Both paths share the question, model parameters, and ground-truth answer. CE provides answer supervision, while KL and attention MSE impose cross-path consistency. Purple squares denote latent tokens, while blue squares denote text tokens.

ABSTRACT

Learning visual latent states that effectively support answer generation remains challenging for multimodal large language models (MLLMs). Existing approaches often rely on explicit visual supervision or staged restrictions on image access, introducing dependencies on predefined visual targets or multi-stage training. We introduce **CrystaL**, a single-stage dual-path framework that learns task-relevant visual latent states through their utility for answer generation. An intact path derives latent states from the original image, while a parameter-shared corrupted path leverages the transferred states to predict the same answer under degraded visual input. Joint answer supervision and cross-path consistency encourage the latent states to preserve information useful for prediction when direct visual evidence is weakened. CrystaL requires neither external visual feature targets nor intermediate visual annotations, and retains only the intact path at inference. Across nine benchmarks, CrystaL achieves relative gains of 6.0% and 4.6% over Qwen2.5-VL-7B and Qwen3-VL-4B-Instruct, respectively, supporting the effectiveness of the proposed dual-path training framework in learning visual latents that support answer generation.

1 INTRODUCTION

Visual reasoning requires multimodal large language models (MLLMs) to identify and use the visual evidence relevant to a question. Latent reasoning complements textual chain-of-thought (Zhang et al., 2023; Shao et al., 2024) by introducing intermediate hidden states that can carry information without explicitly verbalizing it (Chen et al., 2025; Gao et al., 2025; Yang et al., 2025). A central challenge is

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.

therefore to learn latent states that retain visual information useful for answer generation. However, answer supervision alone does not specify what these intermediate states should preserve. Moreover, when the decoder retains direct access to the image, it may predict the answer without substantially relying on the latent states. The resulting problem is not merely how to introduce latent computation, but how to make that computation functionally useful for visual reasoning.

Existing approaches address this problem through two broad strategies. One line of work explicitly guides what the latent states should encode. Aurora (Bigverdi et al., 2025) and CoVT (Qin et al., 2025), for example, obtain supervision from external vision experts, while LVR (Li et al., 2025a), SKILA (Tong et al., 2025), and Monet (Wang et al., 2025b) introduce auxiliary visual reasoning signals or intermediate visual data. These methods make the latent states visually informative by providing explicit targets for their content. A second strategy instead increases the model's dependence on latent computation by restricting direct visual access. LIVR (Li et al., 2025b), for instance, first trains under a visual bottleneck and subsequently restores full image access. These strategies either prescribe what latent states should represent or rely on staged changes in visual information access. The former may limit their adaptability to diverse visual reasoning tasks, while the latter may weaken latent utilization once full visual access is restored.

Table 1: Method properties. Training requirements and reasoning space.

Property	Aurora	SKILA	LVR	LIVR	CoVT	Ours
No extra module	✗	✗	✓	✓	✗	✓
No extra images	✓	✗	✓	✓	✓	✓
One-stage training	✗	✓	✗	✗	✗	✓
Continuous reasoning	✗	✓	✓	✓	✓	✓

This raises a different question: *Can visual latent states become useful for answer generation without prescribing their content and without changing visual access across training stages?* Our key observation is that latent learning can be viewed as a problem of *information routing*. Rather than directly specifying what the latent states must imitate, we can construct a prediction condition in which answer generation has greater need for the information carried by those states. At the same time, the states themselves should still be produced from complete visual evidence. This suggests separating the source of the latent states from the auxiliary prediction condition used to train it.

We instantiate this idea with **CrystaL**, a single-stage dual-path framework. An *intact path* processes the original image and produces latent states from complete visual evidence. A parameter-shared *corrupted path* receives a degraded view of the same image and uses the intact-conditioned latent states to answer the same question. Because direct visual evidence is weakened only in the corrupted path, successful prediction places greater functional demand on the transferred latent states, while their source path continues to observe the full image. The two paths are optimized jointly with answer supervision and cross-path consistency, requiring neither external feature teachers nor intermediate visual annotations. At inference, the corrupted path is removed and only the intact path is retained. In this way, visual degradation serves only as a training mechanism for learning useful latent states rather than as a restriction on test-time perception.

We evaluate CrystaL across nine benchmarks covering spatial reasoning, high-resolution perception, visual search, hallucination, and general visual understanding. On Qwen2.5-VL-7B, CrystaL improves the aggregate score by 6.0% relative to the base model and outperforms matched Direct-FT across all main evaluation metrics. On Qwen3-VL-4B-Instruct, it yields a further 4.6% relative aggregate improvement, indicating that the gains persist across two generations and scales of Qwen-VL backbones. Together, these results support the effectiveness of the proposed single-stage training formulation for learning visual latent states that benefit answer generation.

Our contributions are threefold:

- **Learning principle.** We formulate visual-latent learning as an information-routing problem: instead of prescribing what latent states should encode, we create a prediction condition in which answer generation has greater need for information carried by latent states.
- **Method.** We introduce CrystaL, a single-stage dual-path framework that learns task-relevant visual latent states by transferring intact-conditioned representations to a visually degraded prediction path, jointly optimizing answer supervision and cross-path consistency to encourage the preservation of information useful for answer generation.

- **Empirical validation.** Experiments across nine benchmarks and two Qwen-VL backbones demonstrate the effectiveness of CrystaL in learning visual latent states that support answer generation, achieving relative aggregate gains of 6.0% and 4.6%, respectively.

2 Related Work

2.1 Multimodal Chain-of-Thought Reasoning

Inspired by the success of CoT approaches in pure language models, recent research has focused on endowing MLLMs with similar multi-step reasoning capabilities. Early text-based CoT methods employed prompts to guide models in generating explicit reasoning steps before providing final answers(Zhang et al., 2023; 2025; Xu et al., 2025). Although effective for scientific Q&A and mathematical problems(Lu et al., 2023), these approaches remain constrained by the limited expressiveness of discrete linguistic symbols when describing continuous visual details. Models tend to rely on self-generated text rather than raw visual inputs, leading to information loss and hallucination(Wu et al., 2025b; Dong et al., 2025).

To address this, tool-augmented frameworks such as DeepEyes(Zheng et al., 2025) and Visual Programming(Gupta & Kembhavi, 2023) delegate perception tasks to external vision models. However, such pipelines are non-differentiable and heavily reliant on the quality of off-the-shelf tools. Recent hybrid approaches attempt to achieve visually anchored reasoning by interleaving visual features with text(Su et al., 2025), but these typically require costly human annotation of reasoning trajectories or bounding box supervision. In contrast, our proposed CrystaL framework cultivates an intrinsic reasoning capability that is fully differentiable, eliminating reliance on external tools or dense annotations.

2.2 Latent Space Reasoning

Moving beyond explicit text generation, research has shifted toward latent reasoning, in which inference unfolds through a sequence of hidden states. In language models, Coconut(Hao et al., 2024) and Quiet-STaR(Zelikman et al., 2024) show that models can perform intermediate computation through implicit vectors before producing an answer, without verbose textual reasoning. This paradigm is now extending to multimodal large language models. Mint-CoT(Chen et al., 2025) maps patch indices to visual tokens for direct supervision. Monet(Wang et al., 2025b) aligns intermediate token features with images that depict reasoning trajectories, while CoVT(Qin et al., 2025) uses external visual models to supervise latent visual features. Although these methods provide explicit targets for latent content, their objectives do not directly test whether the resulting states remain useful for prediction when visual evidence is weakened. CrystaL instead optimizes latent utility through cross-path consistency under corrupted visual input.

2.3 Self-Supervised Visual Representation Learning

Our training paradigm draws inspiration from self-supervised learning in computer vision. Masked image modeling(He et al., 2022) and contrastive learning(Chen et al., 2020; Grill et al., 2020) show that robust visual representations can be learned by reconstructing corrupted inputs or enforcing consistency across views. Although originally designed for pretraining conventional encoders, these principles have recently been applied to post-training alignment in multimodal large language models(Li et al., 2025c). For example, Visual Jigsaw(Wu et al., 2025a) enhances fine-grained perception through a jigsaw puzzle task, and CapPO(Tu et al., 2025) reduces hallucination through caption consistency. These approaches typically operate at the output level. CrystaL applies consistency regularization directly within the latent reasoning space, requiring states derived from intact images to support prediction under corrupted visual input.

3 Method

The introduction motivates a training condition with two requirements: latent states should be formed from complete visual evidence, yet answer prediction should have a reason to use them. CrystaL satisfies both requirements within a single-stage, parameter-shared dual-path framework

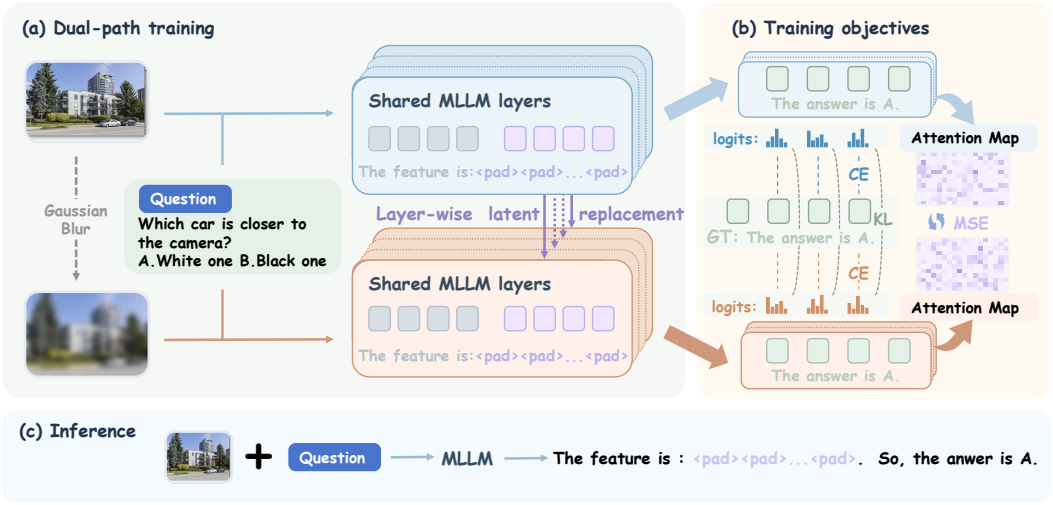


Figure 2: Overview of CrystaL. (a) Parameter-shared intact and corrupted paths receive the same question with original and degraded images, respectively. Intact-conditioned latent states are transferred to the corrupted path at each decoder layer. (b) Answer supervision and cross-path consistency jointly optimize latent generation and utilization across the two visual conditions. (c) At inference, only the intact path is retained to generate the latent tokens followed by the answer.

(Fig. 2). The intact path generates latent states from the original image. The corrupted path receives a degraded image and predicts the same answer using latent states transferred from the intact path. Joint supervision trains the model to generate these states and to use them for prediction. At inference, only the intact path is retained.

3.1 Task and Latent Representation

Given an image I, a question Q, and an answer $A = (a_1, \ldots, a_M)$, we place a segment of K latent tokens $Z = (z_1, \ldots, z_K)$ before the answer. The model generates $Y = (Z, A)$ autoregressively:

$$p_\theta(Y \mid I, Q) = \prod_{t=1}^{K+M} p_\theta(y_t \mid I, Q, y_{<t}). \tag{1}$$

The tokens mark positions for latent computation; the representations of interest are their continuous hidden states. Let Ω_Z and Ω_A denote the latent and answer positions, respectively. For path $b \in \{\text{int}, \text{cor}\}$, write $\mathbf{H}_b^{(\ell)}[\Omega_Z] \in \mathbb{R}^{K \times d}$ for the latent-position states at decoder layer ℓ. We use teacher forcing during training and generate both Z and A autoregressively at inference.

3.2 Dual-Path Information Routing

The two paths assign different roles to the same training example. The intact path supplies the latent states from complete visual evidence; the corrupted path tests whether those states can support prediction when direct visual evidence is weakened. Both paths share parameters, the question, and the answer target. Their visual inputs differ: $I_{\text{int}} = I, I_{\text{cor}} = \mathcal{C}(I; \rho)$, where $\mathcal{C}$ is an image-corruption operator with strength ρ. We use Gaussian blur by default and examine other operators and strengths in Appendix A.2. Corruption affects the auxiliary prediction path; it does not remove visual evidence from the path that generates the latent states.

Maintaining both visual conditions throughout training serves two complementary purposes. The intact path preserves the visual context used at inference, while the corrupted path introduces an auxiliary prediction task that encourages the transferred states to retain answer-relevant information. Parameter sharing couples these roles within the same model: the model trained to answer with reduced direct visual evidence is also the model used to generate latent states from the original image. This coupling maintains the auxiliary learning signal throughout training without requiring a transition between restricted and restored visual access.

Transferring intact-conditioned states. Degrading the second path creates a need for information, but does not by itself provide a route for that information. We establish this route by transferring the intact path's latent-position states at every decoder layer. At layer ℓ, let $\mathbf{T}_{\text{lat}}^{(\ell)} = \mathbf{H}_{\text{int}}^{(\ell)}[\Omega_Z]$. Because the latent segment precedes the answer in causal decoding, these states cannot attend to future ground-truth answer tokens. We then replace the corresponding states in the corrupted path without detachment:

$$\widetilde{\mathbf{H}}_{\text{cor}}^{(\ell)}[j] = \begin{cases} \mathbf{T}_{\text{lat}}^{(\ell)}[k], & j = \Omega_Z[k], \\ \mathbf{H}_{\text{cor}}^{(\ell)}[j], & j \notin \Omega_Z, \end{cases} \qquad \ell = 1, \ldots, L. \tag{2}$$

Image and non-latent text positions retain their corrupted-path states. Thus, the auxiliary answer predictor receives a degraded visual context together with latent states formed from the intact image. It receives neither intact-image token states nor intact-path answer states. The transfer makes the latent positions the designated route by which intact-conditioned information reaches the corrupted-path prediction.

3.3 Joint Training Objectives

Making intact-conditioned states available does not guarantee that the model will use them. We therefore train prediction under both visual conditions and align the paths at two levels: their answer distributions and their answer-to-latent attention. Let $\mathbf{p}_{b,t}$ be the predictive distribution at output position t in path b.

Supervised generation. The intact path learns to generate the latent segment and the answer, whereas the corrupted path is supervised on the answer alone:

$$\begin{aligned} \mathcal{L}_{\text{ce}}^{\text{int}} &= -\frac{1}{K+M} \sum_{t=1}^{K+M} \log \mathbf{p}_{\text{int},t}(y_t), \\ \mathcal{L}_{\text{ce}}^{\text{cor}} &= -\frac{1}{M} \sum_{m=1}^{M} \log \mathbf{p}_{\text{cor},K+m}(a_m). \end{aligned} \tag{3}$$

The first term trains generation with full visual context. The second trains answer prediction with weaker direct visual evidence and access to transferred latent states. Together, they connect latent generation to the downstream task, without specifying an external target for latent content.

Predictive consistency. Both paths predict the same answer, but answer-level supervision alone does not require their token distributions to agree. We additionally align their distributions at corresponding answer positions:

$$\mathcal{L}_{\text{kl}} = \frac{1}{M} \sum_{m=1}^{M} D_{\text{KL}}(\mathbf{p}_{\text{int},K+m} \,\|\, \mathbf{p}_{\text{cor},K+m}) \,. \tag{4}$$

At each position, the paths share the same teacher-forced answer prefix. The KL term therefore asks the corrupted path to maintain the intact path's predictive behavior despite its degraded image.

Attention consistency. Output agreement does not specify how answer tokens interact with the transferred states. We regularize that interaction at the final decoder layer. For each path, average answer-to-latent attention across its H heads:

$$\bar{\mathbf{A}}_b = \frac{1}{H} \sum_{h=1}^{H} \text{Attn}_{b,h}^{(L)}[\Omega_A, \Omega_Z] \in \mathbb{R}^{M \times K}. \tag{5}$$

Let $\sigma(\bar{\mathbf{A}}_b)$ normalize each answer position over the K latent positions. We minimize $\mathcal{L}_{\text{attn}} = \frac{1}{M} \left\| \sigma(\bar{\mathbf{A}}_{\text{int}}) - \sigma(\bar{\mathbf{A}}_{\text{cor}}) \right\|^2$. This term encourages similar relative attention to the latent positions across visual conditions. It complements distributional consistency without assigning semantic labels to individual latent tokens.

Overall objective and inference. The complete single-stage objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}}^{\text{int}} + \mathcal{L}_{\text{ce}}^{\text{cor}} + \mathcal{L}_{\text{kl}} + \mathcal{L}_{\text{attn}}$. The objectives constrain complementary aspects of the prediction task. Cross-entropy anchors both paths to the ground-truth sequence; predictive consistency aligns their answer distributions; attention consistency regularizes the relative allocation of answer-to-latent attention. Together, these signals guide latent learning through downstream prediction and cross-path agreement, while leaving the visual content of individual latent states unspecified.

At inference, the auxiliary corrupted path and cross-path objectives are removed. Given the original image and question, the intact path autoregressively generates the latents followed by the answer.

4 Experiments

4.1 Experimental Setup

The main evaluation covers CVBench (Tong et al., 2024a), HRBench (Wang et al., 2025c), BLINK (Fu et al., 2024), RealWorldQA (Zhang et al., 2024), V*Bench (Wu & Xie, 2024), and POPE (Li et al., 2023), yielding eight reported metrics. We also test MMBench (Liu et al., 2023), MMStar (Chen et al., 2024), and MMVP (Tong et al., 2024b), and extend the evaluation to the Qwen3-VL-4B-Instruct (Bai et al., 2025) backbone.

Baselines. We compare with the base model, CoVT(Qin et al., 2025), LVR(Li et al., 2025a), LIVR(Li et al., 2025b), SKILA(Tong et al., 2025), Vision-R1(Huang et al., 2025), Monet(Wang et al., 2025b), PixelReasoner(Wang et al., 2025a), and VThinker(Qiao et al., 2025). Public checkpoints evaluated locally are marked with † in Table 2. Claude Sonnet 4(Anthropic, 2025) and GPT-4o(OpenAI, 2024) provide closed-source reference points.

Table 2: Main comparison. Accuracy (%) on eight evaluation metrics. Average is the unweighted mean over all eight metrics and is reported for complete rows. Bold and underline denote the best and second-best open-source results in each column, respectively; – denotes unavailable results. † denotes a public checkpoint evaluated locally.

Method	CVBench		HRBench		BLINK	RWQA	V*	POPE	Average
	2D	3D	4K	8K					
Closed-source reference models									
Claude-4-Sonnet	76.3	79.0	32.3	22.7	39.6	63.7	15.2	80.1	51.1
GPT-4o	55.0	55.2	50.6	46.7	63.0	69.7	42.9	87.0	58.8
Open-source models									
Qwen2.5-VL-7B	75.0	73.3	68.6	64.9	55.7	68.6	76.4	86.5	71.1
CoVT†	75.1	**84.7**	71.0	<u>69.4</u>	56.0	**71.6**	78.0	84.6	73.8
LVR	–	78.7	69.6	66.1	52.5	67.7	81.7	–	–
Vision-R1	8.0	37.6	64.8	57.0	51.0	9.4	80.1	77.8	48.2
LIVR	73.9	82.7	<u>72.4</u>	69.0	<u>56.3</u>	69.2	79.1	**89.5**	<u>74.0</u>
SKILA	–	–	72.0	66.5	**56.7**	70.5	**84.3**	87.3	–
PixelReasoner†	<u>75.6</u>	78.5	69.9	65.3	54.6	68.9	77.0	87.0	72.1
VThinker†	68.5	55.3	59.6	59.9	32.1	66.0	75.4	69.3	60.8
Monet	–	73.3	71.0	68.0	52.9	69.3	<u>83.3</u>	–	–
CrystaL	**76.6**	<u>84.4</u>	**73.4**	**71.1**	55.9	<u>70.6</u>	82.7	<u>88.7</u>	**75.4**
Δ vs. base	+1.6	+11.1	+4.8	+6.2	+0.2	+2.0	+6.3	+2.2	+4.3

4.2 Main Results

Gains with intact visual input. CrystaL improves the base model on every metric in the main evaluation suite (Table 2). The gains are particularly pronounced on CVBench-3D, high-resolution perception, and visual search, while remaining positive on general visual understanding and hallucination evaluation. This pattern suggests that the learned latent pathway is most beneficial when prediction requires integrating spatially distributed or fine-grained visual evidence. Crucially, these

gains are measured using only the intact path at inference, showing that the corrupted path acts as an effective training constraint rather than a test-time crutch.

Beyond task-specific fine-tuning. CrystaL consistently outperforms the matched Direct-FT control across all eight metrics, with the clearest separation on visual search (Fig. 3). This uniform advantage highlights the benefit of routing supervision through the dual-path latent-learning framework beyond conventional task-specific fine-tuning.

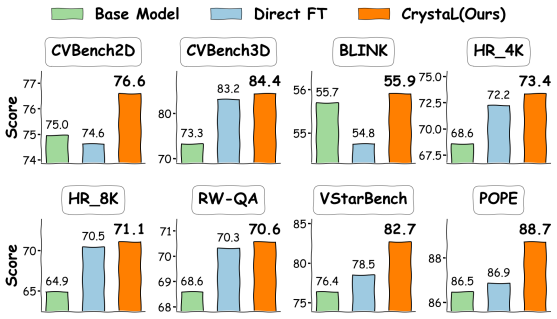


Figure 3: Matched Direct-FT control. Accuracy (%) on the eight main metrics. Direct-FT and CrystaL use the same backbone and task data.

Comparison with prior methods. CrystaL achieves the strongest overall result among open-source methods with complete evaluations. It combines leading performance on spatial and high-resolution benchmarks with strong results on visual search, general visual understanding, and hallucination evaluation. The resulting performance profile demonstrates that CrystaL learns broadly useful visual latents without external visual teachers or staged training.

4.3 Generalization

Benchmark transfer. CrystaL transfers effectively across both evaluation suites and backbone configurations (Tables 3 and 4). It improves the base model on all three additional benchmarks, with particularly strong gains on MMVP, and improves five of six metrics on Qwen3-VL-4B-Instruct, with the clearest gains in high-resolution perception. Together, these results demonstrate stable transfer across datasets, model generations, and scales.

Table 3: Dataset generalization. Accuracy (%) on Qwen2.5-VL-7B; Δ denotes CrystaL's gain over the base in percentage points. Bold and underline denote the best and second-best results, respectively.

Method	MMB	MMStar	MMVP
Qwen2.5-VL	82.5	60.3	66.0
CoVT	<u>84.5</u>	**65.1**	**78.3**
LVR-7B	74.6	61.3	72.0
SKILA	83.3	<u>64.8</u>	75.3
CrystaL	**85.0**	63.5	<u>77.0</u>
Δ vs. base	+2.5	+3.2	+11.0

Table 4: Backbone generalization. Accuracy (%) with Qwen3-VL-4B-Instruct. AVG is the unweighted mean over the six metrics; Δ is the change in percentage points.

Metric	Base	+ Ours	Δ
BLINK	52.3	55.3	+3.0
CV-2D	70.4	75.2	+4.8
CV-3D	90.3	89.5	−0.8
HR-4K	65.5	70.1	+4.6
HR-8K	60.6	66.0	+5.4
RWQA	67.3	68.9	+1.6
AVG	67.7	70.8	+3.1

4.4 Mechanistic Analysis

We directly examine the predictive contribution of the transferred states by intervening on latent access while varying visual quality and answer-option order (Table 5). Across these controlled conditions, intact-conditioned latent states consistently strengthen the corrupted path and provide reusable information for answer generation.

Latent utility under degraded input. Supplying intact-conditioned states improves performance for every reported CVBench subset under both clean and blurred inputs. Under blur, the states consistently recover part of the accuracy lost when direct visual evidence is degraded; under clean input, they provide an additional benefit rather than becoming redundant. The shared direction of the effect across visual conditions supports the interpretation that the latent states carry task-relevant information available to downstream prediction.

Table 5: Controlled latent intervention. CVBench accuracy (%). Latent states come from the clean-image path; Blur + S shuffles the recipient's options. ADE/COCO are 2D source subsets.

Input	Latent	2D ADE	2D COCO	2D All	3D All
Clean	✗	66.1	77.7	72.6	80.7
Clean	✓	**67.9**	**82.2**	**75.9**	**84.7**
Blur	✗	54.9	67.5	62.0	74.1
Blur	✓	**58.0**	**68.6**	**63.9**	**76.6**
Blur + S	✗	54.4	65.1	60.5	73.6
Blur + S	✓	**57.5**	**67.7**	**63.2**	**76.1**

Option-mapping control. The benefit remains after answer options are shuffled, even though the supplied states are generated before shuffling and kept fixed. Their continued effectiveness under a changed answer mapping shows that the latent states capture reusable task-relevant information beyond fixed option identities. Together, the clean, degraded, and shuffled-option results provide strong behavioral evidence that CrystaL learns latents that actively support downstream prediction.

4.5 Ablation Studies

Having established the predictive utility of the transferred states, we next identify the design choices that maximize this benefit. The ablations highlight three clear principles: distinct latent positions promote richer representations, cross-path objectives provide complementary supervision, and structure-preserving corruption creates the most effective visual bottleneck. All ablations use Qwen2.5-VL-7B and the same six evaluation metrics, with results averaged over repeated runs.

Table 6: Latent-token identity and count. Accuracy (%) for repeated or distinct token identities. Base has no latent fine-tuning; * marks the default. AVG covers six metrics. Bold marks the best configuration per metric across both panels.

K	CV 2D	CV 3D	BLINK	HR 4K	HR 8K	RWQA	AVG
(a) Repeated identity							
Base	75.0	73.3	55.7	68.6	64.9	68.6	67.7
4	74.7	74.7	52.9	73.8	71.0	68.0	69.2
8	75.7	83.7	54.4	**74.4**	**72.0**	69.3	71.6
16	72.3	80.6	52.3	71.5	68.1	69.7	69.1
(b) Distinct identities							
4	75.1	82.9	55.3	73.8	70.8	69.8	71.3
8*	**76.6**	**84.4**	**55.9**	73.4	71.1	70.6	**72.0**
16	74.0	72.6	53.9	74.3	71.4	**71.6**	69.6

Table 7: Loss and alignment strategy. Accuracy (%). (a) Intact-path CE is retained; All adds KL and Attn to corrupted-path CE. (b) All/Lat selects attention keys; Soft/Raw denotes MSE after/before softmax. Bold marks each panel's best.

Setting	CV 2D	CV 3D	BLINK	HR 4K	HR 8K	RWQA	AVG
(a) Loss components							
CE	74.7	83.3	54.6	73.0	70.3	69.9	71.0
+ KL	75.2	83.6	54.1	**75.0**	**71.1**	**71.1**	71.7
+ Attn	76.2	84.0	**57.1**	72.6	70.4	70.3	71.8
All	**76.6**	**84.4**	55.9	73.4	**71.1**	70.6	**72.0**
(b) Alignment strategy							
All/Soft	73.6	79.5	53.9	**74.8**	69.4	69.5	70.1
Lat/Raw	76.3	83.1	54.0	73.8	**71.4**	70.2	71.4
Lat/Soft	**76.6**	**84.4**	**55.9**	73.4	71.1	**70.6**	**72.0**

Latent-token identity and count. Distinct token identities yield a higher average than a repeated identity at every tested latent length (Table 6). Both identity schemes achieve their strongest overall performance with an intermediate latent budget, identifying eight tokens as the most effective configuration. The combination of distinct identities and an eight-token sequence provides the best balance between representational diversity and compact latent computation.

Loss components and alignment strategy. Both cross-path objectives improve the aggregate result over corrupted-path answer supervision alone, and their combination performs best overall (Table 7(a)). The benchmark-level variation is informative: output consistency and attention consistency sometimes favor different tasks, suggesting that they regularize complementary aspects of the prediction process rather than duplicating the same signal.

The alignment formulation further shows that where and how consistency is imposed matters (Table 7(b)). Restricting alignment to answer-to-latent attention is more effective than matching attention over all keys, and normalizing the attention values before alignment performs better than matching raw values in this comparison. The latent-only, softmax-normalized alignment directly reinforces the information route from latent states to answer generation and achieves the strongest performance.

Corruption operator. Gaussian blur provides the strongest aggregate performance among the five corruption operators, with random masking as the closest alternative (Table 8). By suppressing local detail while retaining global scene structure, Gaussian blur creates an effective condition for learning latents that complement the remaining visual evidence. The additional strength study in Appendix A.2 identifies moderate blur as the most effective setting, confirming the importance of a structure-preserving visual bottleneck.

Table 8: Training-time corruption. Intact-path accuracy (%) after training with each operator. Mask/Jig. denote masking/jigsaw. AVG covers six metrics; bold marks each metric's best.

Operator	CV 2D	CV 3D	BLINK	HR 4K	HR 8K	RWQA	AVG
Blur	**76.6**	84.4	**55.9**	**73.4**	**71.1**	**70.6**	**72.0**
Mask	76.3	**84.7**	55.2	71.9	69.4	68.4	71.0
Jig.	73.6	82.5	54.0	71.4	69.0	65.2	69.3
Noise	74.9	81.8	54.1	72.0	69.8	69.9	70.4
Color	74.8	80.3	54.6	72.3	70.4	**70.6**	70.5

4.6 Qualitative Analysis

We finally inspect the internal organization of the learned representations. CrystaL exhibits lower off-diagonal similarity across latent positions than CoVT tokens derived from the same visual teacher, whereas CoVT tokens from different teachers form visible teacher-specific blocks (Figure 4). This structure is consistent with the token-identity ablation: distinct latent positions support more differentiated representations even without prescribed feature targets. Together with the controlled interventions, the visualization shows that CrystaL develops both differentiated internal representations and strong functional utility without predefined visual supervision. Appendix A.3 further illustrates the resulting improvements across counting, depth, relation, and distance tasks.

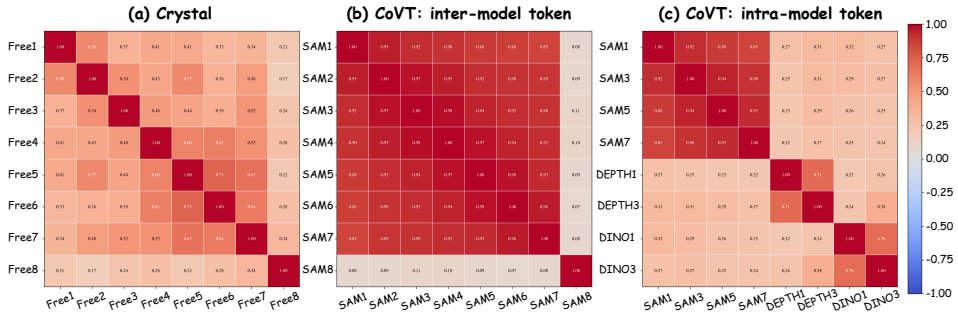


Figure 4: Token-representation structure. Pairwise cosine similarities between (a) eight CrystaL latent tokens, (b) CoVT tokens from the same external teacher, and (c) CoVT tokens grouped by SAM, depth, and DINO supervision. Diagonal entries are self-similarities; color denotes cosine similarity on the shared scale.

5 Conclusion

We introduced CrystaL, a single-stage dual-path framework that learns useful visual latents without external visual teachers or intermediate visual annotations. Across nine benchmarks and two Qwen-VL backbones, CrystaL consistently improves the base model and achieves the strongest average among complete open-source evaluations on the main suite. Controlled latent interventions further show that intact-conditioned states supply reusable, answer-relevant information to a visually weakened path. These results establish dual-path information routing as an effective approach to learning visual latents for answer generation.

AI USE DISCLOSURE

The authors used OpenAI's ChatGPT and Codex to assist with language editing, restructuring and polishing the experimental discussion, summarizing high-level patterns already present in author-provided results, and LaTeX table formatting. These tools were not used to generate experimental data, conduct experiments, develop the proposed method, retrieve literature, or prove mathematical claims. The authors checked all AI-assisted text against the underlying results and take full responsibility for the final manuscript.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 8076–8084, 2019.

Anthropic. System card: Claude opus 4 & claude sonnet 4. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3836–3845, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. URL https://arxiv.org/abs/2403.20330.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. Mirage: Assessing hallucination in multimodal reasoning chains of mllm. *arXiv preprint arXiv:2505.24238*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19520–19529, 2025.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14953–14962, 2023.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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

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

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. *arXiv preprint arXiv:2512.21218*, 2025b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Yuxuan Li, Yicheng Zhang, Wenhao Tang, Yimian Dai, Ming-Ming Cheng, Xiang Li, and Jian Yang. Visual instruction pretraining for domain-specific foundation models. *arXiv preprint arXiv:2509.17562*, 2025c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. URL https://arxiv.org/abs/2307.06281.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, Peiqing Yang, Shiqiang Lang, Enhui Wan, Xiaowan Wang, Yida Xu, Lan Yang, Chong Sun, Chen Li, and Honggang Zhang. V-Thinker: Interactive thinking with images. *arXiv preprint arXiv:2511.04460*, 2025. URL https://arxiv.org/abs/2511.04460.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, et al. Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms. *arXiv preprint arXiv:2510.01954*, 2025.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024b. URL https://arxiv.org/abs/2401.06209.

Songjun Tu, Qichao Zhang, Jingbo Sun, Yuqian Fu, Linjing Li, Xiangyuan Lan, Dongmei Jiang, Yaowei Wang, and Dongbin Zhao. Perception-consistency multimodal large language models reasoning via caption-regularized policy optimization. *arXiv preprint arXiv:2509.21854*, 2025.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a. URL https://arxiv.org/abs/2505.15966.

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

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

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

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms. *arXiv preprint arXiv:2509.25190*, 2025a.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025b.

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

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

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D Goodman. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

Jieyu Zhang, Le Xue, Linxin Song, Jun Wang, Weikai Huang, Manli Shu, An Yan, Zixian Ma, Juan Carlos Niebles, Silvio Savarese, et al. Provision: Programmatically scaling vision-centric instruction data for multimodal language models. *arXiv preprint arXiv:2412.07012*, 2024.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

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

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

A Appendix

A.1 Experiment Details

Training datasets. We use the training suite of CoVT(Qin et al., 2025) without modification: vision-centric subsets of LLaVA-OneVision(Li et al., 2024), TallyQA(Acharya et al., 2019), and ADE20K-Depth(Zhou et al., 2017) in the configuration reported by that work.

Training configuration. We fine-tune with LoRA(Hu et al., 2022) using rank $r = 16$, scaling $\alpha = 32$, and learning rate 2×10^{-4} on four NVIDIA A40 GPUs.

Baseline settings. SKILA(Tong et al., 2025) uses the reported Unified Reasoning Model rather than SKILA-V; LVR(Li et al., 2025a) scores are those reported in the SKILA paper. We locally evaluate the CoVT-seg-dino-depth checkpoint(Qin et al., 2025) under its official setup. For LIVR(Li et al., 2025b), we follow the reported two-stage protocol with 4K steps per stage.

Additional-result provenance. PixelReasoner and VThinker results use publicly available checkpoints evaluated with VLMEvalKit. The Monet row contains the available overlapping main-suite scores reported during rebuttal; its remaining cells are unavailable. The additional-benchmark table uses the consolidated evaluation results, without combining them with earlier revisions. The Qwen3-VL-4B-Instruct extension is reported separately from Qwen2.5-VL-7B.

A.2 Additional Ablations

Section 4.5 examines latent-token configuration, loss components, attention-alignment strategy, and the five individual corruption operators. Here we test alignment depth, mixed corruptions, and corruption strength to further characterize the training conditions.

Alignment depth. The topmost decoder layer obtains the best average, demonstrating that answer-to-latent alignment is most effective near the model output (Table 9). This result complements the layer-wise latent replacement by identifying the most effective depth for the attention objective.

Mixed corruptions. Gaussian blur achieves the strongest aggregate result among the single and mixed corruption strategies (Table 10). Its consistent advantage highlights the value of using a stable, structure-preserving corruption signal during latent learning.

Table 9: Alignment depth. Accuracy (%) when attention alignment is applied at one decoder layer. Depth is counted from the output side of the 28-layer decoder; Top, 7th, 14th, and 21st are separate single-layer settings, not cumulative layer blocks.

Metric	Top	7th	14th	21st
BLINK	55.9	53.3	54.8	49.6
CV-2D	76.6	71.2	75.9	74.7
CV-3D	84.4	83.2	84.7	82.5
HR-4K	73.4	74.4	73.6	74.0
HR-8K	71.1	70.6	69.6	70.9
RWQA	70.6	72.0	68.8	68.1
AVG	72.0	70.8	71.2	70.0

Table 10: Mixed-corruption strategy. Accuracy (%) for Gaussian blur, random masking (Cut), or an equal-probability mixture of the two (Mix). Blur and Cut repeat the single-operator results.

Metric	Blur	Cut	Mix
BLINK	55.9	55.2	55.0
CV-2D	76.6	76.3	75.7
CV-3D	84.4	84.7	84.9
HR-4K	73.4	71.9	74.0
HR-8K	71.1	69.4	71.9
RWQA	70.6	68.4	68.5
AVG	72.0	71.0	71.7

Training-time corruption strength. Moderate blur, sampled from $\mathrm{Uni}(2, 5)$, achieves the strongest aggregate result across the tested corruption strengths (Tables 11 and 12). This setting provides the best balance between suppressing direct visual detail and retaining scene structure, establishing moderate degradation as the preferred training configuration.

Table 11: Training-time blur strength. Accuracy (%) with no corruption or Gaussian-blur radii sampled uniformly from the displayed ranges. AVG is the unweighted mean over six metrics.

Metric	None	2–5	5–10
BLINK	54.1	55.9	55.3
CV-2D	76.6	76.6	77.3
CV-3D	85.0	84.4	82.2
HR-4K	71.0	73.4	72.1
HR-8K	68.6	71.1	69.0
RWQA	69.9	70.6	71.9
AVG	70.9	72.0	71.3

Table 12: Strong training-time corruption. Accuracy (%) under moderate blur, strong blur, or a fully black corrupted-path image. The 2–5 setting is repeated from the left panel for direct comparison.

Metric	2–5	10–20	Black
BLINK	55.9	54.1	53.0
CV-2D	76.6	75.9	74.4
CV-3D	84.4	81.0	83.5
HR-4K	73.4	73.5	73.5
HR-8K	71.1	70.4	70.8
RWQA	70.6	69.0	69.8
AVG	72.0	70.7	70.8

A.3 Qualitative Examples Across Tasks

Figures 5 and 6 highlight CrystaL's successful reasoning across depth, counting, relation, and distance tasks, illustrating the breadth of its visual improvements.

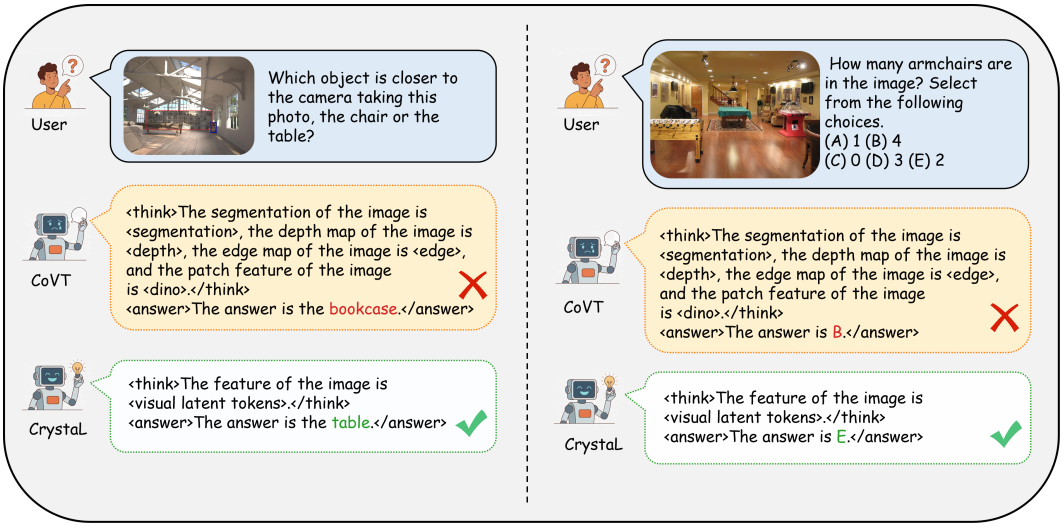


Figure 5: Selected depth and counting examples. The upper row shows the image and question; the lower rows compare CoVT and CrystaL outputs. CrystaL correctly identifies the nearer object in the depth example (left) and the number of armchairs in the counting example (right), whereas CoVT answers both incorrectly.

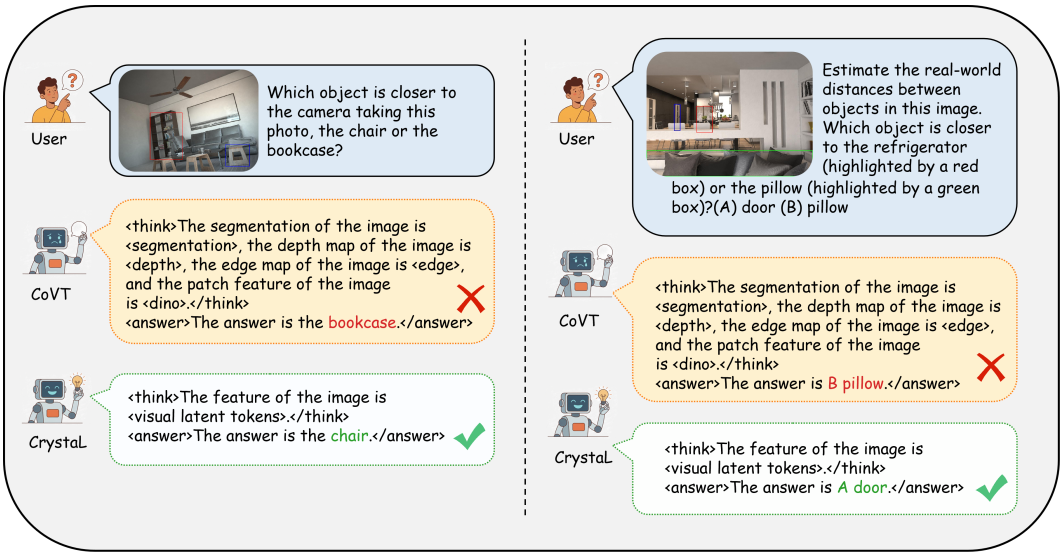


Figure 6: Selected relation and distance examples. The upper row shows the image and question; the lower rows compare CoVT and CrystaL outputs. CrystaL correctly resolves the relative-depth query (left) and the object-to-refrigerator distance query (right), whereas CoVT answers both incorrectly.

A.4 Visualizing Corruption Operators

Figures 7–11 provide visual references for the five corruption operators used in the ablations.



Figure 7: Gaussian blur. Original image and smoothing at $r = 2, 5, 10$.

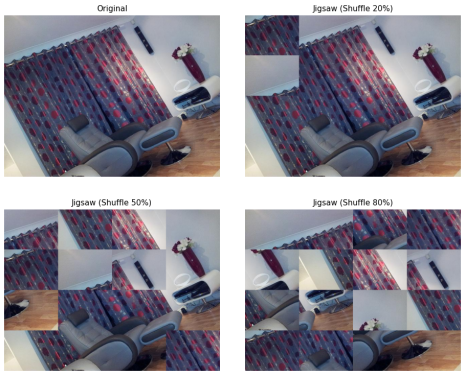


Figure 8: Jigsaw. Shuffling ratios $\rho = 20\%, 50\%, 80\%$.

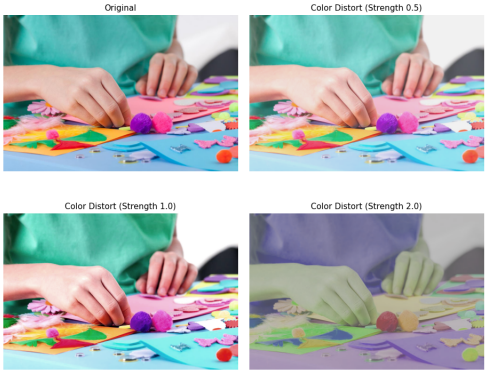


Figure 9: Color distortion. Color-jitter scales $s = 0.5, 1.0, 2.0$.

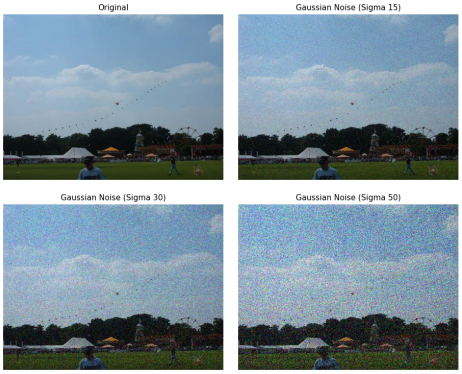


Figure 10: Gaussian noise. Intensities $\sigma = 15, 30, 50$.

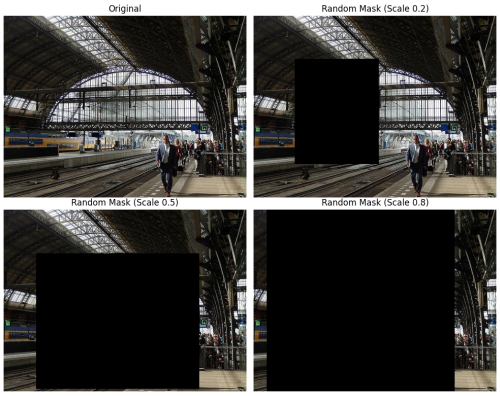


Figure 11: Random masking. Original image and erasing scales $s = 0.2, 0.5, 0.8$.

A.5 Attention Maps During Training

The following examples qualitatively compare the two paths' attention patterns; they are illustrative and are not used as quantitative evidence of latent content.

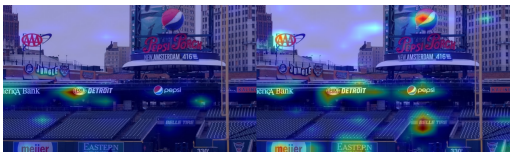


Figure 12: Label-focused attention. Intact-path attention (left) concentrates on "Detroit," whereas corrupted-path attention (right) is distributed across the labels.



Figure 13: Person-focused attention. Intact-path attention (left) is scattered, whereas corrupted-path attention (right) concentrates on the man.



Figure 14: Text-focused attention. Intact-path attention (left) is scattered, whereas corrupted-path attention (right) concentrates on the text behind the man.