COUNTERFACTUAL INTERVENTIONS FOR LATENT VISUAL REASONING

Anonymous authors
Paper under double-blind review

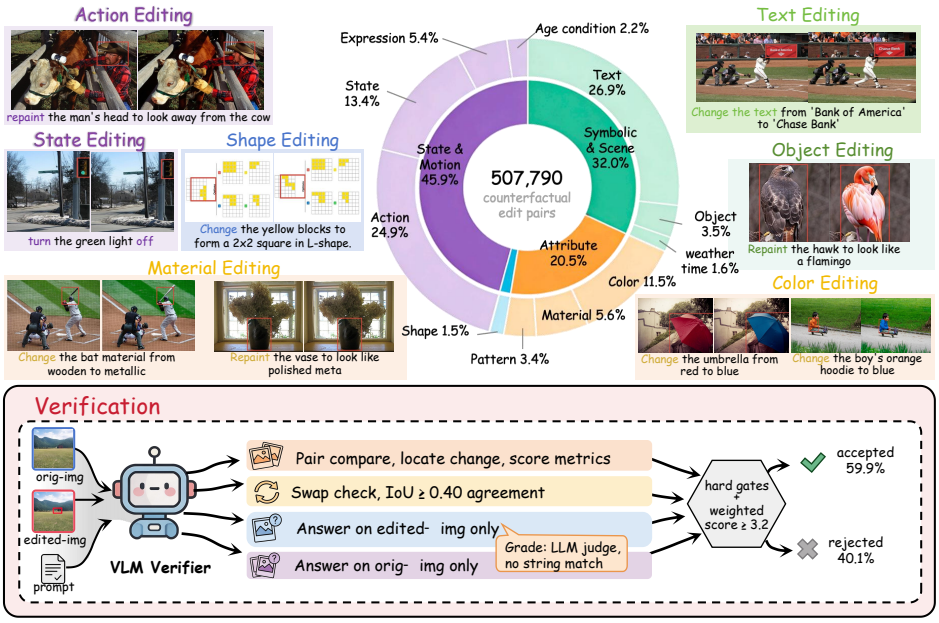


Figure 1: How we produce and verify the counterfactual pairs.

ABSTRACT

Latent visual reasoning offers intermediate computation between direct answering and explicit textual chain-of-thought. Its value depends on whether the answer uses the evidence carried by continuous states. Answer supervision and representation alignment can leave this dependence weak: a decoder may answer from the image while bypassing the latent. We observe that counterfactual image pairs address this gap by providing both a meaningful replacement state and the answer that should follow its transfer. Unlike arbitrary perturbations, evidence-changing edits supply known targets for latent interventions. Continuous substitution passes answer gradients through the transferred state, while paired vector differences provide directions for exploring evidence-related variations. We introduce **CILVR**, which uses this relation in supervised latent learning and policy optimization. Supervised training combines alignment with differentiable swaps targeting the donor image's answer. Its reinforcement stage, **CA-GRPO**, concentrates exploration along paired latent differences and rewards correctness and consistency under transfer. An automated editing and verification pipeline supplies **CFVR**, a dataset of verified counterfactual image pairs with reasoning traces on both sides. **CILVR-8B** achieves competitive performance across ten vision-centric and reasoning benchmarks. Ablations compare paired data, training objectives, and sensitivity to latent interventions. Code and data will be released upon acceptance.

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.

1 INTRODUCTION

Latent visual reasoning processes visual evidence through continuous intermediate states. Recent methods reconstruct visual features, compress multimodal reasoning traces, and propagate information through latent steps (Li et al., 2025a; Yang et al., 2025; Wang et al., 2025a). These states offer a compact workspace between direct answering and explicit textual chain-of-thought (Wei et al., 2022; Hao et al., 2025; Chen et al., 2025): the model performs intermediate computation without expressing every step in language. Direct answering can suffice when the relevant evidence is immediately accessible. Latent computation is useful when the model must retain or combine observations before deciding. Its value therefore depends on whether that additional computation changes how the answer uses visual evidence.

Existing objectives constrain outputs or intermediate content. Answer supervision rewards correct responses, while alignment encourages latents to match visual or reasoning targets (Li et al., 2025a; Wei et al., 2025). Neither directly specifies how an answer should change when latent evidence changes. A decoder may rely on the image even when its latents match their targets. We call this potential failure mode *latent bypass*. Analyses of weak input-to-latent and latent-to-answer connections (Li et al., 2026) motivate training latent use alongside answer correctness.

Our observation is that counterfactual pairs fit continuous latent spaces by supplying interventions with explicit answer targets for otherwise opaque states. Textual chain-of-thought exposes readable steps for supervision; continuous latents lack this semantic interface. Replacing an arbitrary latent therefore leaves the desired answer unspecified. A counterfactual image pair changes answer-relevant evidence while keeping the question fixed, providing both a meaningful replacement state and its answer. Holding the receiving image and question fixed, we can transfer the other image's latent and train toward its answer. For example, an orange-to-blue shirt edit supplies a latent whose transfer should shift the answer to "blue", while ordinary supervision retains "orange" for the original image without a swap. Counterfactual pairs also support direct-answer and textual reasoning supervision (Xu et al., 2025). Continuous states offer a distinct opportunity: a transferred text trace is a sequence of discrete tokens, whereas latent substitution preserves a differentiable path to the state generator. Answer gradients can therefore train both the production and use of transferred evidence, without requiring a textual interpretation of each intermediate state. Their vector differences also define directions for exploring variations associated with the evidence change. The same pair thus specifies how latent interventions should affect answers and which latent variations are worth exploring.

We introduce **CILVR**, which organizes supervised learning and policy optimization around this relation. Supervised training combines alignment with differentiable swaps that target the donor answer. **CA-GRPO** then samples continuous trajectories with exploration concentrated along paired latent differences. Correctness and consistency under transfer provide group-relative rewards (Shao et al., 2024) for latent generation and answer decoding. An automated editing and verification pipeline constructs **CFVR**, supplying controlled evidence changes and paired answer targets (Figure 1). Experiments compare downstream accuracy, training objectives, and sensitivity to latent interventions.

Our contributions are:

- **Counterfactual training for latent reasoning.** We propose CILVR: latent swaps supervise donor-answer prediction, while paired latent differences and transfer rewards guide policy exploration.
- **Counterfactual data construction.** Our editing and verification pipeline builds CFVR, with 304,302 verified image pairs sharing a question but requiring different answers, with reasoning traces on both sides.
- **Evaluation across reasoning formats.** Experiments on ten benchmarks compare direct answering, text CoT, and latent reasoning, with ablations of training objectives and latent interventions.

2 Related Work

Thinking with images. Explicit visual reasoning uses search, grounding, and image operations. V* performs guided search (Wu & Xie, 2024); Visual CoT grounds reasoning in regions (Shao et al., 2025); CogCoM and ReFocus use manipulations and image edits (Qi et al., 2024; Fu et al., 2025); and GRIT supervises bounding-box grounding (Fan et al., 2025). Reinforcement learning extends these operations: DeepEyes learns zooming (Zheng et al., 2025), Pixel Reasoner acts in pixel space (Su et al., 2025), and Chain-of-Focus adapts visual search (Zhang et al., 2025b). Perception-aware rewards and verifiable-reward training further improve multimodal reasoning (Xiao et al., 2026; Cao et al., 2025; Zhang et al., 2025a; Shen et al., 2025). Explicit steps remain interpretable but incur decoding or tool costs.

Latent-space reasoning. Coconut feeds hidden states back as embeddings (Hao et al., 2025), Soft Thinking mixes concept tokens (Zhang et al., 2026b), and SIM-CoT supervises implicit steps (Wei et al., 2025). For visual reasoning, LVR reconstructs query-relevant features (Li et al., 2025a), Mirage generates compressed visual imagery (Yang et al., 2025), Monet distills interleaved reasoning (Wang et al., 2025a), and CrystaL aligns intact and corrupted image paths (Zhang et al., 2026a). LIVR uses a learnable latent bottleneck (Li et al., 2025b). RL variants optimize hidden-state or soft-token policies (Yue et al., 2025b; Zheng et al., 2026; Deng et al., 2026); DeepLatent uses Gaussian modulations (Lu et al., 2026), and CoLVR perturbs latent trajectories for contrastive rewards (Ding et al., 2026). These approaches motivate continuous exploration, while analyses highlight weak connections between latents and answers (Li et al., 2026). CILVR extends counterfactual image pairs (Niu et al., 2021; Zhang et al., 2024a) to latent replacement targets and evidence-related exploration.

3 Method

CILVR trains compact intermediate computation around the relation between visual evidence, latent states, and answers. A counterfactual pair provides two complementary signals: a replacement latent with a known answer target for supervision, and a direction of latent variation for policy exploration. Both stages reward answers that follow the evidence carried by the latent. Figure 2 summarizes supervised training; Section 3.4 describes pair construction.

3.1 Framework Overview

A pair $\tau = (I, I', q, a, a', b)$ contains an image, a local edit within box b, a shared question, and distinct answers $a \neq a'$. The model produces $z = f_\theta(I, q)$ and $z' = f_\theta(I', q)$, then answers through $p_\theta(a \mid I, q, z)$. Between `<lat>` and `</lat>`, each top-layer hidden state is fed back as the next input embedding. The resulting sequence $z = (z_1, \ldots, z_K)$ adds continuous computation before response decoding. The intermediate latent states are passed directly between steps as continuous vectors. Training combines correct answers on each image with latent transfer: we fix the receiving image and question, substitute the other side's latent, and target the donor answer. The image path remains available. The transfer objective therefore trains the decoder to respond to latent evidence even when the receiving image supports a different answer. Correctness on the unmodified inputs anchors this behavior to the visual task.

3.2 Counterfactual Latent Supervision

Preparing latent content. Warm-up on 279K pairs teaches the latent format through next-token supervision on written reasoning and answers, yielding θ_0. Self-distillation then supplies intermediate targets: a teacher pass uses the same network without gradient and takes the hidden state at the last token of each reasoning step as its target. The student rolls out its own latents and aligns one state per step by cosine similarity over the last 16 layers. Both passes see the image; only the teacher sees the written reasoning. Each side of a pair has its own targets, preserving their distinct visual content.

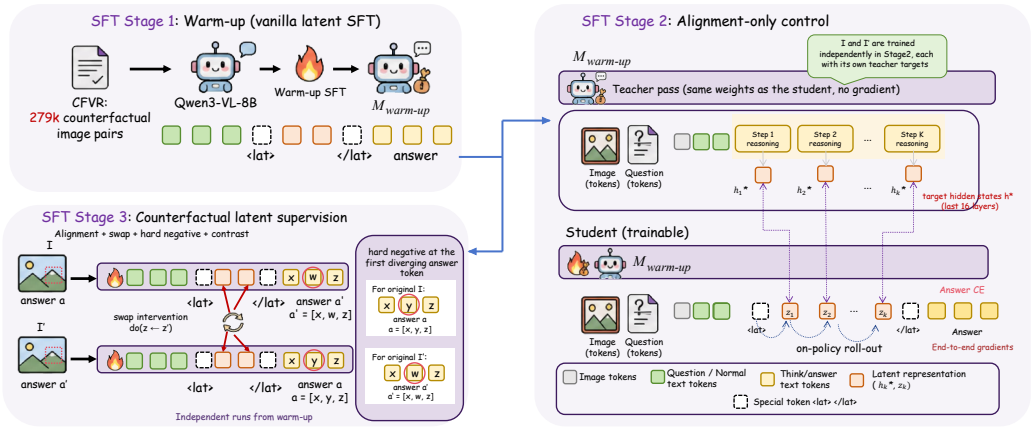


Figure 2: Supervised components: warm-up, self-distillation, and latent intervention. The alignment-only control and the full intervention objective both start from the warm-up checkpoint; Stage 3 does not continue from the Stage 2 control.

Training latent use. Alignment specifies what the states should contain. To constrain how they affect the answer, we exchange the paired latent sequences and optimize

$$\mathcal{L}_{\text{swap}} = -\log p_\theta(a' \mid I, q, z') \; - \; \log p_\theta(a \mid I', q, z). \tag{1}$$

For example, if an edit changes a shirt from orange to blue, transferring z' into the original image context targets "blue"; the reverse transfer targets "orange". The pair supplies both a meaningful substitute and its target answer. Continuous substitution lets answer gradients train the production and use of the transferred states. Ordinary paired answer supervision constrains the two outputs separately; the swap loss additionally couples each output to the other side's latent.

The full objective retains normal answer supervision and alignment. A stopping term teaches span termination, a hard negative separates the answers at their first diverging token, and cross-pair contrast encourages distinct latent representations. Stage 2 is the alignment-only control; Stage 3 jointly optimizes alignment and intervention from the same θ_0, then initializes RL. Appendix A gives the auxiliary losses and schedule.

3.3 Counterfactual-Anchored Policy Optimization

Supervised swaps establish a target for latent use. **CA-GRPO** further searches for latent trajectories that support both correct answers and successful transfer. The paired difference guides where to explore, while the transfer reward evaluates whether the sampled states retain the answer-relevant distinction.

Direction and policy. Under the current policy, we deterministically generate an anchor span on each image and mean-pool its positions into $\bar{\mu}$ and $\bar{\mu}'$. Their normalized nonzero difference defines

$$u = \frac{\bar{\mu}' - \bar{\mu}}{\|\bar{\mu}' - \bar{\mu}\|}. \tag{2}$$

The direction is recomputed as the policy changes and held fixed without gradient within a pair. It is associated with the edit, but may also encode answer identity or span length. At step k, the policy samples

$$z_k \sim \pi_\theta(\cdot \mid I, q, z_{<k}) = \mathcal{N}\big(\mu_k,\ \Sigma_k\big), \qquad \Sigma_k = \|\mu_k\|^2 \Big[s_\parallel^2\, uu^\top + \frac{s_\perp^2}{d-1} \left(\mathbf{I} - uu^\top\right) \Big], \tag{3}$$

where μ_k is the hidden state given the sampled prefix, d is its dimension, and the norm is treated as constant. The two scales allocate noise along u and across its orthogonal subspace; division by $d-1$ keeps the latter budget from growing with dimension. A larger parallel budget concentrates

exploration along the paired difference, while orthogonal noise preserves variation beyond that direction. Samples are fed back autoregressively, so a perturbation can affect subsequent latent states. We draw G trajectories per image, each containing a latent span and a decoded answer.

Groups and rewards. CA-GRPO follows GRPO (Shao et al., 2024): the critic is replaced by a baseline computed from a group of rollouts that share the same input. The two images of a pair are different inputs with different reference answers. Each pair therefore yields two groups, $\mathcal{G}_I$ and $\mathcal{G}_{I'}$, of G rollouts each. For rollout i, $c_i \in \{0, 1\}$ scores its own answer. We transfer its span to the other image and decode greedily; f_i is one if the output matches the donor side's reference answer. Thus RL evaluates the same relation trained by $\mathcal{L}_{\text{swap}}$:

$$r_i = c_i + \beta\, f_i, \qquad A_i = \frac{r_i - \text{mean}_{j \in \mathcal{G}(i)}\, r_j}{\text{std}_{j \in \mathcal{G}(i)}\, r_j + \epsilon}, \tag{4}$$

where $\mathcal{G}(i)$ is the group containing rollout i. Rewards are compared only among rollouts of the same image, so A_i compares trajectories for a fixed input. A group with identical rewards gives zero advantage. The two groups stay coupled through the shared direction u and through f_i, which scores every span on the other image.

Objective. For sampled answer a_i and latent span z_i of length K_i, the trajectory score is

$$\ell_i(\theta) = \frac{1}{|a_i|} \log p_\theta(a_i \mid I_i, q, z_i) + \lambda_z \frac{1}{K_i\, d} \sum_{k=1}^{K_i} \log \mathcal{N}\big(z_{i,k};\, \mu_k, \Sigma_k\big). \tag{5}$$

The two terms normalize answer log-probability by length and latent log-density by span length and dimension. The latent term explicitly scores the sampled continuous states, allowing the same reward to optimize both latent generation and answer decoding. With the score ratio $\rho_i(\theta) = \exp\big(\ell_i(\theta) - \ell_i(\theta_{\text{old}})\big)$, we maximize

$$\mathcal{J}(\theta) = \mathbb{E}_\tau\Big[\sum_{s \in \{I, I'\}} \frac{w_s}{2G} \sum_{i \in \mathcal{G}_s} \Big(\min\big(\rho_i A_i,\ \text{clip}(\rho_i, 1-\varepsilon, 1+\varepsilon)\, A_i\big) - \beta_{\text{KL}}\, D_i \Big) \Big], \tag{6}$$

$$D_i = \exp\big(\ell_i(\theta_{\text{ref}}) - \ell_i(\theta)\big) - \big(\ell_i(\theta_{\text{ref}}) - \ell_i(\theta)\big) - 1. \tag{7}$$

Here $w_I = 1$, and a smaller $w_{I'}$ accounts for imperfect edits. The reference policy θ_{ref} is the Stage 3 model that initializes RL. Each batch of rollouts is used for μ gradient updates. D_i penalizes deviations from the reference in normalized trajectory scores. No critic is used.

Latent interventions. A counterfactual transfer specifies a target answer. Evidence-agnostic replacements instead supply an auxiliary penalty on correct answers that persist after intervention. For each image we decode under three intervened spans: an all-zero span, a span of random vectors with matched norms, and a span generated for another question. Let $\tilde{c}_m \in \{0, 1\}$ indicate that the greedy answer under intervention m is still correct. These rollouts stay outside the group statistics and enter an auxiliary clipped surrogate with the fixed advantage $\tilde{A}_m = -\gamma\, \tilde{c}_m$ and are scored by their length-normalized answer log-probability. The term vanishes once an intervention changes the answer. This term is added to the group objective.

3.4 Counterfactual Pair Construction

The training objectives require pairs with a shared question, a targeted evidence change, and distinct answers. Our pipeline first selects an editable region and designs a question around its visible properties, then edits the image and verifies the pair (Figures 1 and 3). Minimizing changes to non-target content limits unrelated variation in both the latent swap and the paired difference.

Step 1: selecting a region and designing the question. We start from image–question–answer triplets drawn from natural scenes in Visual-CoT (Shao et al., 2025) and reasoning images in Thyme-RL, DeepEyes, and other collections (Zhang et al., 2025c; Zheng et al., 2025). Qwen3.5-397B (Team, 2026) first selects a small box b containing a visible property that can be edited. It then designs a question q, its original answer a, and an edit instruction with a different target answer a'. The source question and answer serve as context and may be reused, adapted, or replaced. The designed question stays fixed across the pair. Edits cover objects, text, states, colors, actions, shapes, and materials; the intended answer change is specified before editing.

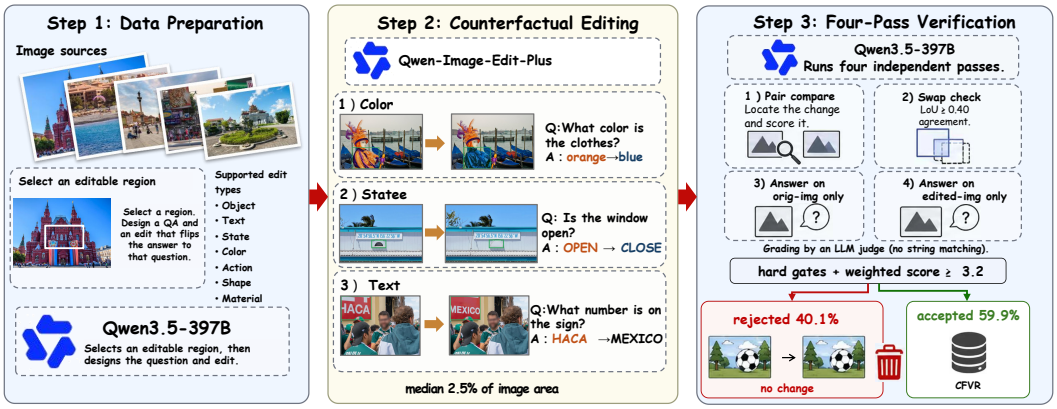


Figure 3: Selecting editable evidence, designing a shared question, and verifying the image edit and paired answers.

Step 2: editing the region. Qwen-Image-Edit-Plus uses the instruction and target region b to produce I', aiming to preserve non-target content. The median box area is 2.5% of the image. For example, a shirt can change from orange to blue while retaining the surrounding scene content. Verification allows mild global color or brightness drift; weather and time-of-day edits may also affect lighting beyond the target region. Locality is therefore assessed by the extent of unrelated content changes, without requiring pixel-wise invariance.

Step 3: four-pass verification. An edit may ignore the instruction, alter unrelated content, damage the image, or leave the answer unchanged. We check each candidate with Qwen3.5-397B in four separate passes. The pair pass sees I, I', and the instruction, and scores edit success, locality, and fidelity. A blind localization pass finds the changed region without receiving its location; the predicted box must overlap b with an IoU of at least 0.4. Two answer passes solve q independently on I and I', and their outputs must match a and a'. An LLM judge checks answer equivalence to accommodate paraphrases. A candidate passes verification only if it clears every hard gate and its weighted score over the four passes reaches 3.2. These checks assess whether the pair supplies the intended evidence change and reliable targets for latent transfer.

Dataset scale and training subsets. **CFVR** comprises a generated pool of 507,790 edited and rendered pairs. The verified subset with reasoning traces on both sides contains 304,302 pairs. These counts distinguish generated pairs from verified pairs with traces; individual training stages use the subsets described in Section 4.1. Appendix B provides the construction and verification prompts, and Appendix C reports the dataset breakdown and example edits.

4 Experiments

4.1 Setup

Training and evaluation. We train rank-64 LoRA adapters on Qwen3-VL-8B-Instruct (Bai et al., 2025a), freeze its vision tower, and cap images at 8 megapixels. After warm-up, Stage 2 and Stage 3 independently run 1.5k steps on 42K pairs with teacher traces. CA-GRPO starts from Stage 3 and runs 400 updates on a 6K-pair pool, with eight pairs per batch and $G = 8$ rollouts per image. Appendix A gives the full schedule and hyperparameters. At test time, the model receives the images and question, emits up to eight latents, and decodes greedily without a prompt requesting textual CoT. For our runs, a pool of twelve LLM judges grades semantic equivalence with the reference. Appendix A.4 documents evaluation provenance.

Benchmarks and baselines. We evaluate on ten benchmarks with 13,673 questions. MMVP (Tong et al., 2024), V* (Wu & Xie, 2024), and HR-Bench 4K and 8K (Wang et al., 2025b) target fine-grained perception. BLINK (Fu et al., 2024) tests multi-image comparison, and

Method	Fine-grained perception				General and multi-image			
	MMVP	V*	HR-4K	HR-8K	MMStar	BLINK	MMBench	RWQA
Proprietary models								
GPT-4o (Hurst et al., 2024)	70.7	66.0	59.0	55.5	61.6	60.0	82.1	68.6
GPT-4o-mini (Hurst et al., 2024)	–	–	–	–	54.8	53.6	77.6	67.1
GPT-5	87.7	72.8	74.6	72.0	76.4	71.0	–	82.8
Gemini 3.1 Pro	–	89.0	90.8	86.4	–	81.2	–	79.3
Claude Sonnet 4.5	74.7	–	–	–	74.1	64.8	–	61.1
Open-source models								
Gemma3 27B (Team et al., 2025)	–	–	–	–	59.6	53.5	82.3	62.5
Vision-R1 7B (Huang et al., 2025)	72.6	80.1	64.8	57.0	60.7	51.0	79.8	67.1
Qwen2.5-VL 7B (Bai et al., 2025b)	66.0	78.5	68.5	65.0	60.3	55.1	82.5	67.4
Qwen3-VL-8B (Bai et al., 2025a)	78.3	83.8	76.5	74.3	**67.4**	64.4	89.4	67.2
+ SFT	75.7	<u>84.8</u>	77.9	<u>75.9</u>	65.3	64.6	88.1	67.2
+ GRPO	79.0	82.7	75.2	71.9	64.5	**66.4**	89.5	70.6
Latent reasoning models								
LVR 7B (Li et al., 2025a)	72.0	81.7	69.6	66.1	61.3	52.5	74.6	67.7
Monet 7B (Wang et al., 2025a)	–	83.3	71.0	68.0	–	–	–	–
SkiLa-V (Tong et al., 2025)	72.0	79.6	68.5	65.2	64.5	56.7	82.4	70.5
SkiLa (Tong et al., 2025)	75.3	84.3	72.0	66.5	64.8	56.7	83.3	69.3
CILVR Stage 1	81.3	81.2	74.0	68.6	63.2	60.4	88.8	67.5
CILVR Stage 2	<u>83.3</u>	81.7	76.9	72.0	63.4	59.4	89.4	69.4
CILVR Stage 3	82.3	81.7	**78.8**	74.0	63.9	61.0	<u>90.0</u>	<u>71.4</u>
CILVR-8B (ours)	**84.1**	**86.4**	<u>78.0</u>	**76.4**	<u>67.0</u>	<u>66.3</u>	**90.6**	**71.9**
Δ vs. Qwen3-VL-8B	+5.8	+2.6	+1.5	+2.1	−0.4	+1.9	+1.2	+4.7

Table 1: Accuracy (%) on fine-grained perception and general multimodal benchmarks. Bold: best and underline: second best among open-source models. Shaded rows: CILVR-8B and its gain over the backbone. Evaluation details appear in Appendix A.4.

every model receives all images of a question. MMStar (Chen et al., 2024), MMBench (Liu et al., 2024), and RealWorldQA (xAI, 2024) cover general visual question answering. MME-RealWorld (Zhang et al., 2024b) covers high-resolution scenes, reported as perception and reasoning. VisualPuzzles (Song et al., 2025) isolates reasoning from domain knowledge, with five categories: algorithmic, analogical, deductive, inductive, and spatial. We report accuracy per benchmark and the macro average. Proprietary baselines are GPT-4o and GPT-4o-mini (Hurst et al., 2024), GPT-5, Claude Sonnet 4.5, Gemini 2.5 Pro, and Gemini 3.1 Pro. Appendix A.4 gives the sources and settings for the added reference results. Open-source generalists are Gemma3 27B (Team et al., 2025), Vision-R1 7B (Huang et al., 2025), DeepEyes (Zheng et al., 2025), Pangea 7B (Yue et al., 2025a), Qwen2.5-VL 7B (Bai et al., 2025b), and our backbone Qwen3-VL-8B. Latent reasoning models are LVR (Li et al., 2025a), Monet (Wang et al., 2025a), and SkiLa and SkiLa-V (Tong et al., 2025). Two controls train the backbone on the same counterfactual pairs with text chain-of-thought, by SFT and by GRPO.

4.2 Main Results

Table 1 reports the vision-centric results. CILVR-8B improves over Qwen3-VL-8B on seven of eight benchmarks, with its largest gains on MMVP and RealWorldQA, and leads the external latent reasoning baselines in all eight columns. CILVR-8B is competitive with the text-CoT controls while using continuous intermediate states for reasoning.

The supervised controls distinguish alignment from intervention. Stage 2 and Stage 3 start from the same warm-up checkpoint; the full intervention objective improves six of eight columns over alignment alone. The CA-GRPO results improve seven columns over Stage 3, with the largest difference on BLINK.

Method	Perception	Reasoning	Overall
Proprietary models			
GPT-4o	54.4	48.3	52.0
GPT-4o-mini	38.8	35.2	37.4
GPT-5	–	–	57.8
Open-source models			
Vision-R1 7B	49.8	44.9	47.9
Qwen2.5-VL 7B	48.6	37.7	44.3
Qwen3-VL-8B	51.0	42.5	47.7
Latent reasoning models			
LVR 7B	52.7	43.5	49.1
Monet 7B	58.3	51.1	55.5
SkiLa-V	51.7	47.7	50.2
SkiLa	56.6	50.2	54.1
CILVR Stage 1	49.7	41.2	46.4
CILVR Stage 2	54.3	40.2	48.9
CILVR Stage 3	62.2	**53.9**	58.9
CILVR-8B (ours)	**63.2**	53.5	**59.4**
△ vs. backbone	+12.2	+11.0	+11.7

Table 2: Accuracy (%) on MME-RealWorld-Lite. Bold: best and underline: second best among open-source models; shaded: ours.

Method	Overall	Alg.	Ana.	Ded.	Ind.	Spa.
Proprietary models						
Gemini 2.5 Pro	49.5	60.0	64.0	60.0	29.7	36.4
Gemini 3.1 Pro	73.5	–	–	–	–	–
GPT-5	57.3	–	–	–	–	–
GPT-4o	41.3	49.2	58.3	49.0	27.3	26.2
Open-source models						
DeepEyes	33.0	37.8	27.0	41.0	26.8	27.0
Pangea 7B	31.3	32.4	23.7	38.5	**28.7**	32.5
Qwen2.5-VL 7B	32.7	37.0	21.8	**47.5**	26.3	21.8
Qwen3-VL-8B	30.1	37.0	38.9	39.0	13.4	23.1
Latent reasoning models						
LVR 7B	29.4	31.3	25.6	40.5	24.4	26.2
Monet 7B	35.0	**45.8**	30.8	**47.5**	26.8	25.5
CILVR Stage 1	32.2	30.5	32.5	41.7	27.6	30.4
CILVR Stage 2	33.5	35.5	29.1	45.8	25.6	32.2
CILVR Stage 3	33.2	37.0	27.0	46.0	21.5	**33.9**
CILVR-8B (ours)	**35.9**	38.2	**39.8**	**47.5**	23.0	32.2
△ vs. backbone	+5.8	+1.2	+0.9	+8.5	+9.6	+9.1

Table 3: Accuracy (%) on VisualPuzzles by reasoning category. Bold: best and underline: second best among open-source models; shaded: ours.

Training data	Algorithm	MMVP	V*	HR_{4K}	HR_{8K}	MMStar	BLINK	VisPuz	MMB	RWQA	MME-RW	Avg.
Qwen3-VL-8B, no training		78.3	83.8	76.5	74.3	**67.4**	64.4	30.1	89.4	67.2	47.7	67.9
No CoT												
SFT data	SFT	79.7	81.2	73.5	70.2	62.6	64.3	34.3	90.0	68.5	46.6	67.1
SFT data	GRPO	81.7	84.3	75.0	71.5	63.4	65.4	34.9	89.8	69.8	49.4	68.5
SFT + RL data	SFT	79.3	81.2	74.1	71.1	62.8	63.5	33.7	89.6	67.6	47.3	67.0
SFT + RL data	GRPO	80.7	80.6	75.1	72.5	63.6	**66.4**	33.1	89.9	69.7	46.7	67.8
RL data	SFT	76.0	81.7	73.5	71.1	61.7	59.6	32.3	89.3	68.2	48.0	66.1
RL data	GRPO	80.0	86.4	77.9	74.5	66.1	65.5	34.2	89.7	70.3	49.7	69.4
Text CoT												
SFT data	SFT	75.7	84.8	77.9	75.9	65.3	64.6	30.0	88.1	67.2	50.7	68.0
SFT data	GRPO	79.0	82.7	75.2	71.9	64.5	**66.4**	35.0	89.5	70.6	47.0	68.2
SFT + RL data	SFT	77.0	85.0	77.0	72.2	64.1	64.1	31.6	88.3	70.9	50.3	68.0
SFT + RL data	GRPO	78.0	83.2	76.4	73.1	63.0	65.8	35.3	89.4	70.2	44.1	67.9
RL data	SFT	81.0	83.8	76.8	71.8	65.0	64.1	32.9	88.8	68.6	51.9	68.5
RL data	GRPO	79.0	**89.0**	**80.6**	76.1	67.3	66.1	25.9	89.0	**72.3**	55.8	70.1
Latent												
Originals only	SFT	**85.0**	82.7	75.8	70.9	64.1	60.8	32.7	89.1	69.0	53.4	68.3
Edited only	SFT	64.0	78.5	71.2	66.4	57.4	53.9	30.6	85.6	60.3	28.4	59.6
SFT data	SFT, no interv.	83.3	81.7	76.9	72.0	63.4	59.4	33.5	89.4	69.4	48.9	67.8
SFT data	SFT	82.3	81.7	78.8	74.0	63.9	61.0	33.2	90.0	71.4	58.9	69.5
SFT + RL data	+ CA-GRPO	84.1	86.4	78.0	**76.4**	67.0	66.3	**35.9**	**90.6**	71.9	**59.4**	**71.6**
w/o difference-guided exploration		84.3	81.7	75.3	70.0	63.0	62.1	32.6	89.2	68.5	48.6	67.5
w/o transfer reward		84.7	81.7	75.5	69.9	64.1	61.8	29.6	88.8	68.9	47.9	67.3
w/o latent trajectory scoring		80.0	81.7	77.1	71.1	59.7	60.2	25.6	87.9	70.7	53.7	66.8

Table 4: Accuracy (%) across ten benchmarks, grouped by reasoning format. Original-only and edited-only controls use the corresponding half of the SFT data. The last three rows remove one CA-GRPO component each. Bold/underline: best/second-best per column across all rows, including ties. Avg.: unweighted mean; shaded: CILVR-8B.

Tables 2 and 3 complete the comparison. CILVR-8B reaches 59.4 on MME-RealWorld-Lite and 35.9 on VisualPuzzles, gaining 11.7 and 5.8 over the backbone. It leads the external open-source baselines on both overall scores. The intervention objective contributes most of the MME-RealWorld-Lite gain, while the reported RL row scores higher on VisualPuzzles overall. A gap to proprietary models remains on VisualPuzzles; our evaluation uses an 8B backbone. The category breakdown reports algorithmic, analogical, deductive, inductive, and spatial reasoning separately.

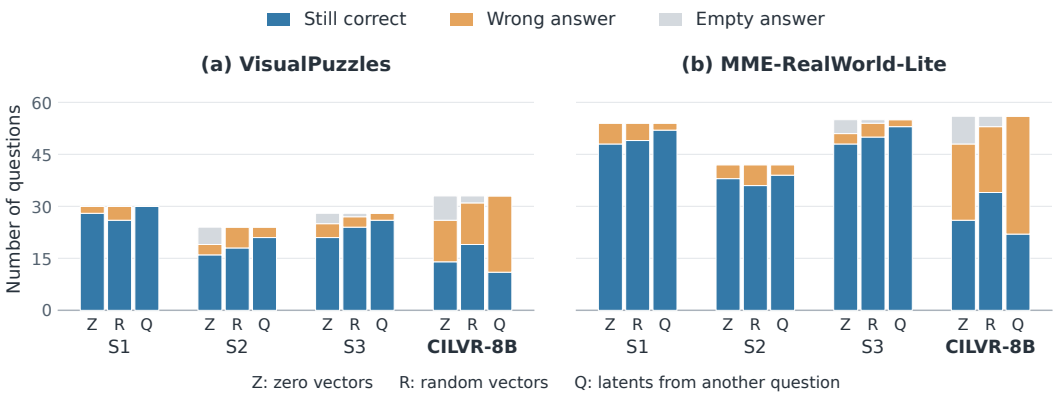


Figure 4: Latent interventions on initially correct answers in VisualPuzzles and MME-RealWorld-Lite. Bars show correct, wrong, and empty outcomes after zeroing (Z), random replacement (R), or other-question replacement (Q). S1–S3 are SFT checkpoints; heights are question counts.

4.3 Training Ablations

Table 4 compares training formats and objectives on paired data. Full latent SFT averages 69.5, compared with 67.8 without intervention. Training on originals alone or edited images alone is also weaker than the full paired objective. The CA-GRPO scores average 71.6, above the strongest direct-answer and text-CoT controls at 69.4 and 70.1.

Removing difference-guided exploration, transfer reward, or latent trajectory scoring yields macro accuracies of 67.5, 67.3, and 66.8, respectively, compared with 71.6 for the full model. These ablations assess the exploration direction, the transfer objective, and the direct optimization of continuous states.

4.4 Latent Ablation

Figure 4 tests each model on questions it initially answers correctly. We replace its latents with zero vectors, random vectors, or another question's latent span. CILVR-8B loses a larger fraction of these correct answers than S1–S3 under all three interventions on both benchmarks.

These interventions remove, perturb, or replace the latent content. Wrong and empty answers are shown separately because they distinguish a changed decision from failed response generation. For CILVR-8B, other-question replacement produces wrong answers without empty outputs. Bar heights reflect each model's initial correct set.

5 Conclusion

We presented CILVR, a framework that uses counterfactual image pairs to train the relation between continuous visual states and answers. Each pair supplies a meaningful replacement latent and the answer that should follow its transfer. Supervised learning uses differentiable swaps to train latent generation and use, while CA-GRPO combines exploration along paired latent differences with rewards for correctness and consistency under transfer. The CFVR pipeline provides the edited and verified training pairs. CILVR-8B improves over its backbone on nine of ten benchmarks, and ablations support combining paired data with intervention objectives. Together, these results motivate training how latent evidence affects the response alongside what the latent represents.

The current study focuses on an 8B backbone and targeted visual edits with a shared question and distinct answers. Further evaluation across model sizes and architectures would establish the scope of this training approach. Future work can extend the framework to questions requiring several coordinated evidence changes, test interventions beyond those used in training, and examine which visual distinctions are retained across latent steps. These settings would assess how latent evidence use generalizes to more complex visual tasks.

AI USE STATEMENT

Generative AI models were used for counterfactual specification, image editing, reasoning-trace generation, automated data verification, answer grading, coding assistance, and writing polish. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and collects no new personal data. All source images and questions come from public datasets used under their original licenses. Local image edits could in principle be misused to fabricate visual evidence, so the released data is intended for research use only.

REPRODUCIBILITY STATEMENT

The paper describes the data construction pipeline, the training objectives, and the evaluation protocol, and lists the hyperparameters. The appendix provides the prompts and the dataset statistics. Code and data will be released upon acceptance.

REFERENCES

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, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Bytedance Seed. Seed1.8 Model Card: Towards Generalized Real-World Agency, 2026. URL https://arxiv.org/abs/2603.20633v3.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *arXiv preprint arXiv:2505.16782*, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding, 2026. URL https://arxiv.org/abs/2601.10611v1.

Jingcheng Deng, Zihao Wei, Liang Pang, Junhong Wu, Shicheng Xu, Zenghao Duan, and Huawei Shen. Latent-grpo: Group relative policy optimization for latent reasoning, 2026. URL https://arxiv.org/abs/2604.27998.

Ziyang Ding, Linjian Meng, Yiming Wu, Yuhan Li, Yuhao Liu, and Zhen Zhao. Colvr: Enhancing exploratory latent visual reasoning via contrastive optimization. *arXiv preprint arXiv:2605.08802*, 2026.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 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 (ECCV)*, 2024.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

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, 2025. URL https://arxiv.org/abs/2412.06769.

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.

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

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.

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

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space, 2026. URL https://arxiv.org/abs/2602.22766.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Dongchen Lu, Zhimo Li, Mao Shu, and Huo Cao. Deeplatent: Think with images via parallel latent visual reasoning. *arXiv preprint arXiv:2606.00562*, 2026.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12700–12710, 2021.

Suhyeong Park, Junha Jung, and Jaewoo Kang. Reason Through the Latent! Making Latent Visual Reasoning Necessary, 2026. URL https://arxiv.org/abs/2609.06746v1.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*, 2024.

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, 2025.

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

Haozhan Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. https://github.com/om-ai-lab/VLM-R1, 2025. Accessed: 2025-02-15.

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

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

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.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

Qixun Wang, Yang Shi, Letian Cheng, Zhuoran Zhang, Yan He, Yuqi Tang, Qi Zhang, Xinlei Yu, Ruizhe Chen, Tianrun Xu, Yuanxing Zhang, Pengfei Wan, Haotian Wang, and Xianghua Ying. Beacon: Knowing When and How to Perform Agentic Visual Reasoning, 2026. URL https://arxiv.org/abs/2607.28595v1.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

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

xAI. Grok-1.5 vision preview. *xAI Blog*, 2024.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward, 2026. URL https://arxiv.org/abs/2506.07218.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. Defacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. *arXiv preprint arXiv:2509.20912*, 2025.

Wenhao Yang, Yu Xia, Jinlong Huang, Shiyin Lu, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Yuchen Zhou, Xiaobo Xia, Yuanyu Wan, Lijun Zhang, and Tat-Seng Chua. Walk the Talk: Bridging the Reasoning-Action Gap for Thinking with Images via Multimodal Agentic Policy Optimization, 2026. URL https://arxiv.org/abs/2604.06777v1.

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.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages, 2025a. URL https://arxiv.org/abs/2410.16153.

Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. Hybrid latent reasoning via reinforcement learning. *arXiv preprint arXiv:2505.18454*, 2025b.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. *arXiv preprint arXiv:2402.13254*, 2024a.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025b.

Yang Zhang, Danyang Li, Yuxuan Li, Xin Zhang, Tianyu Xie, Mingming Cheng, and Xiang Li. Crystal: Spontaneous emergence of visual latents in mllms. *arXiv preprint arXiv:2602.20980*, 2026a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024b.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images, 2025c. URL https://arxiv.org/abs/2508.11630.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *Advances in Neural Information Processing Systems*, 38:168990–169012, 2026b.

Zhi Zheng, Yu Gu, Wei Liu, Yee Whye Teh, and Wee Sun Lee. Soft-grpo: Surpassing discrete-token llm reinforcement learning via gumbel-reparameterized soft-thinking policy optimization, 2026. URL https://arxiv.org/abs/2511.06411.

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.

A Training Details

A.1 Supervised targets and auxiliary losses

Stage 1 minimizes the next-token loss on written reasoning and answers, with latent states produced by the model between the two markers. Training uses the 279K pairs with teacher traces, treating each side as an individual example. The resulting weights θ_0 provide the common initialization for the subsequent supervised objectives. Warm-up teaches the output format but places no direct constraint on latent transfer.

Self-distillation gives each latent a target from the model's own hidden states. A teacher pass runs the same network without gradient on the image, question, and written reasoning steps. Its hidden state at the last token of step k gives target h_k^*. The student receives the image and question, rolls out its own latent sequence, and aligns one latent with each teacher step:

$$\mathcal{L}_{\text{align}} = \frac{1}{16\,K} \sum_{l} \sum_{k=1}^{K} \big[\, 1 - \cos\big(z_k^{(l)},\ h_k^{*(l)}\big) \big], \tag{8}$$

where l ranges over the last 16 layers. Both passes see the image; only the teacher receives the written reasoning. The text head stays frozen, and each side has its own targets. A stop term keeps the span open for K states, while supervision on the closing marker teaches termination. The alignment-only objective, reported as Stage 2, combines the answer loss, $\alpha\mathcal{L}_{\text{align}}$, and the stop term.

Auxiliary constraints and training schedule. A hard negative separates the two answers at their first diverging token j. It encourages the correct continuation to exceed the other answer by margin m:

$$\mathcal{L}_{\text{neg}} = \big[\, m - \log p_\theta(a_j \mid I, q, z, a_{<j}) + \log p_\theta(a'_j \mid I, q, z, a_{<j}) \,\big]_+, \tag{9}$$

with a symmetric term on I'. A cross-pair contrast further encourages separation between the paired latent representations. The full objective adds swap, negative, and contrast terms to the alignment-only objective, weighted by λ_{swap}, λ_{neg}, and λ_{con}. This full run, reported as Stage 3, starts from θ_0 and jointly optimizes alignment and intervention. Stage 2 is a separate control from the same initialization, with the intervention terms removed. The Stage 3 model initializes policy optimization.

A.2 Optimization settings

Stages 2 and 3 each run 1.5k steps on the 42K-pair trace subset, aligning the last 16 layers with $\alpha = 6$. Stage 3 uses $\lambda_{\text{swap}} = 0.5$ and $\lambda_{\text{neg}} = 0.3$. CA-GRPO runs 400 updates over a 6K-pair pool at a learning rate of 3×10^{-6}. Each batch contains eight pairs, with $G = 8$ rollouts per image, and is reused for $\mu = 2$ updates. We set $s_\parallel = 0.1$, $s_\perp = 0.02$, $\lambda_z = 0.1$, $\beta = 0.5$, $\gamma = 1$, $\varepsilon = 0.2$, $\beta_{\text{KL}} = 0.04$, and $w_{I'} = 0.1$.

A.3 Policy score and noise normalization

Equation 5 combines answer and latent terms with weight λ_z. We clip the gradient norm at one. The dimension-normalized covariance gives expected squared displacement $\|\mu_k\|^2(s_\parallel^2 + s_\perp^2)$. With $s_\parallel = 0.1$ and $s_\perp = 0.02$, about 96% of this expectation lies along u.

A.4 Evaluation sources and settings

Backbone and external baseline rows use published scores; trained controls and CILVR use the grading procedure in Section 4.1.

The macro average is the unweighted mean of the ten benchmark accuracies.

Added proprietary references. GPT-5 results on V*, MMStar, BLINK, RealWorldQA, and VisualPuzzles use the *high* setting in Table 2 of Bai et al. (2025a). Its MMVP item accuracy and MME-RealWorld-Lite overall accuracy are from Table 1 of Park et al. (2026), under their CoT evaluation setting. Its HR-Bench scores are from Table 1 of Yang et al. (2026), which reports Avg@8

on both HR splits. Each entry retains its source evaluation setting. Gemini 3.1 Pro scores use Tables 1 and 2 of Wang et al. (2026), whose evaluation uses temperature 0.1 and rule-based matching followed by model-based judgment. Claude Sonnet 4.5 scores on MMVP and MMStar use Table 2 of Bytedance Seed (2026); BLINK validation and RealWorldQA use Table 6 of Clark et al. (2026). Dashes denote unavailable reference scores, including MMBench results without a matching version and split and unreported category scores. Scores are rounded to one decimal place. We do not compute a cross-benchmark average for these partial reference rows.

B PROMPTS USED IN THE DATA PIPELINE

Every prompt below is reproduced verbatim from the source constant that the pipeline actually executes. Braced variables are populated with each sample at call time.

Prompt 1: Counterfactual Edit Specification

```
You are designing a COUNTERFACTUAL visual-reasoning training sample
from a real photo.
Original dataset question (CONTEXT ONLY - you may reuse, adapt, or
replace it):  {question}
Original answer:  {answer}
STEP 1.  Find ONE small region of the image holding a clearly visible
object whose APPEARANCE can be repainted locally by a diffusion image
editor.
STEP 2.  Design a QA whose answer is decided by ONE visible property
of
that region, plus the simplest local edit that clearly FLIPS the
answer.
Report the region as "box":  [x1,y1,x2,y2] in 0-1000 NORMALISED image
coordinates (top-left origin, 0<=x1<x2<=1000, 0<=y1<y2<=1000).  The
box
must be TIGHT around the target object, not the whole image.
Pick the edit_type that fits the region MOST naturally:
- color (car red -> blue) | material (wooden -> metal) | pattern
(striped -> solid)
- text (sign "OPEN" -> "CLOSED") | weather_time (day -> night; box
must cover sky)
- age_condition (clean -> dirty, fresh -> wilted) | state (light on ->
off, door open -> closed)
- action (standing -> sitting, looking left -> right) | expression
(smiling -> frowning)
- object (cat -> dog, same size/shape) | shape (round -> square)
PREFER DIVERSITY: do NOT default to color.  Actively pick
action/expression/state/material/pattern/text/object when the region
supports them.
HARD FORBID (the editor physically CANNOT do these):
- removing OR adding a whole person / object / accessory
- changing the COUNT of objects
- anything needing a redraw of the whole scene or many separate
regions
RULES:
- answer and new_answer must BOTH be SHORT (1-4 words), concrete,
CLEARLY different.
- edit_instruction is a direct imperative repainting ONE appearance
property of ONLY the box content.
- The question must be answerable from the image alone and grounded in
the box region.
- If no region has an appearance property cleanly flippable by a
single local repaint, set editable=false.
Output ONLY one JSON object:
{{"box":[x1,y1,x2,y2],"edit_type":"...","question":"...","answer":"...
","edit_instruction":"...","new_answer":"...","editable":true}}
```

Prompt 2: Step-wise Visual Reasoning Chain

```
You are annotating a visual reasoning chain for training.
Question:  {q}
Correct answer:  {a}
```

```
Produce the SHORT chain of visual inspection steps a careful person
takes to reach that answer.
Each step must name ONE concrete region to look at, and what is read
off it.
Return STRICT JSON only, no prose:
{{"steps":[{{"box":[x1,y1,x2,y2],"think":"..."}}]}}
Rules:
- 2 to 4 steps.  Each "think" is ONE short sentence (<= 25 words),
stating what is observed in that region.
- box coordinates are integers 0-1000, normalized to image
width/height.
- Steps must progress toward the answer; do not restate the question,
do not mention the answer in step 1.
- Ground every step in something visible.  No speculation.
```

Prompt 3: Edit Audit: Execution, Locality, Fidelity

```
You are given TWO images of the same scene:  IMAGE_1 and IMAGE_2.
Exactly one local region was edited between them.
Intended edit:  "{ins}"
Report what you actually observe, not what you expect.  Return STRICT
JSON only:
{{"changed_box":[x1,y1,x2,y2],
"edit_success":1-5,
"locality":1-5,
"fidelity":1-5,
"changed_desc":"..."}}
Definitions:
- changed_box:  the region that differs between the two images,
integers 0-1000
normalized to width/height.  If you see no difference at all, return
[0,0,0,0].
- edit_success:  5 = clearly realized; 4 = realized with minor
deviation;
3 = partially realized but recognizable; 2 = barely; 1 = not
realized.
- locality:  5 = only the target region changed; 4 = target changed,
negligible
elsewhere; 3 = target changed plus a mild global tone/lighting shift
(still the same scene, same layout, same identities); 2 = noticeable
unrelated changes; 1 = whole scene re-rendered.
- fidelity:  5 = natural and artifact-free; 4 = minor softness;
3 = usable with visible but non-distracting artifacts; 1 = garbled.
Be lenient.  These images come from a diffusion editor, so slight
global colour
or brightness drift and mild compression artifacts are EXPECTED and
should not
be heavily penalised.  Judge whether a human reader would still accept
the pair
as "same scene, one thing changed".
- changed_desc:  one short sentence on what actually changed.
```

Prompt 4: Blind Answering for Flip Check

```
Question:  {q}
Answer based only on what is visible in THIS image.
Reply with the answer itself, as briefly as possible.  No explanation.
```

Prompt 5: Semantic Answer Matching

```
A question was asked about two versions of the same image.
Question: {q}
Reference answers:
A = {a}
B = {b}
The model replied:
answer_1 (first image) = "{p1}"
answer_2 (second image) = "{p2}"
For each reply decide which reference it matches SEMANTICALLY.
Synonyms,
paraphrases, different casing, or extra words still count as a match.
If a reply matches neither, or is empty/evasive, answer NEITHER.
Reply STRICT JSON only:
{{"a1":"A"|"B"|"NEITHER","a2":"A"|"B"|"NEITHER"}}
```

Prompt 6 is used for the text-CoT baselines. Latent evaluation supplies the images and question followed by the latent start marker, without this text-CoT instruction.

Prompt 6: Text-CoT Baseline Instruction

```
Look at the image and answer the question.
First reason about the visual evidence step by step inside <think>
</think>, referring to what is actually visible in the relevant
region.
Then output the final answer inside <answer> </answer>. The content
of <answer> must be as short as possible: a word, a phrase, a number,
or a single option letter. Do not repeat the question and do not
explain inside <answer>.
```

Prompt 7: Benchmark Answer Grading

```
You are grading answers to visual questions. For each item decide
whether
the model's answer means the SAME as the reference answer.
Count as correct: synonyms, different wording or casing, numerals vs
words
("12" = "twelve"), extra words that do not change the meaning, a full
sentence
whose core answer matches, or the right option letter.
Count as wrong: a different entity/value, a refusal, an empty answer,
or a
hedge that never commits.
{items}
Reply STRICT JSON only, one verdict per id, no other text:
{{"verdicts":[{{"id":<int>,"correct":true|false}}, ...]}}
```

C Dataset Statistics and Visualization

This section documents the generated data and individual training pairs. Table 5 distinguishes the edited and rendered pool from the verified subset with teacher reasoning on both sides. Table 6 breaks the same pairs down by edit type and by source, which shows that the two image pools contribute different kinds of edits. Figures 5 to 10 then show six pairs drawn from six different edit directions.

Property	Count	Share
Edited and rendered pairs	507,790	100.0%
from VisCoT images	403,587	79.5%
from the RL image pool	104,203	20.5%
Verified pairs with reasoning on both sides	304,302	59.9%
VisCoT subsets	42,260 / 127,571 / 109,212	
RL pool	25,259	
Edit types	12	

Table 5: Scale of the dataset. Every pair is one image with a single edited region and an answer that flips as a result.

Edit type	VisCoT	RL pool	Total
text	114,509	22,189	136,698
action	120,113	6,572	126,685
state	63,207	4,712	67,919
color	16,016	42,383	58,399
material	25,085	3,182	28,267
expression	25,495	1,774	27,269
object	7,146	10,505	17,651
pattern	12,194	5,155	17,349
age_condition	10,169	994	11,163
weather_time	7,995	103	8,098
shape	1,655	6,110	7,765
geometry / chart	3	524	527

Table 6: Edited pairs per edit type. The VisCoT side is dominated by actions and scene text, the RL pool by colour and object attributes. The prompt lists eleven edit types. The annotator wrote a label outside this list for 0.15% of the pairs. We assign each such label to one of the eleven types by its edit instruction. Attributes specific to geometry figures and charts come mostly from the RL pool and form a twelfth type.

Reading a sample page. Each figure places the original image on the left and the edited image on the right, marks the same region on both, and prints the edit instruction together with the answer before and after the edit. Each edit targets a selected visual property and aims to preserve non-target content. Mild global color or lighting changes may remain, particularly for weather and time-of-day edits. The paired answers specify how the response should change with the intended evidence change. The six figures cover six edit directions of increasing difficulty for the editor: colour (Figure 5), action (Figure 6), material (Figure 7), object state (Figure 8), weather and time of day (Figure 9), and surface pattern (Figure 10).



Figure 5: Repainting the colour of one object.

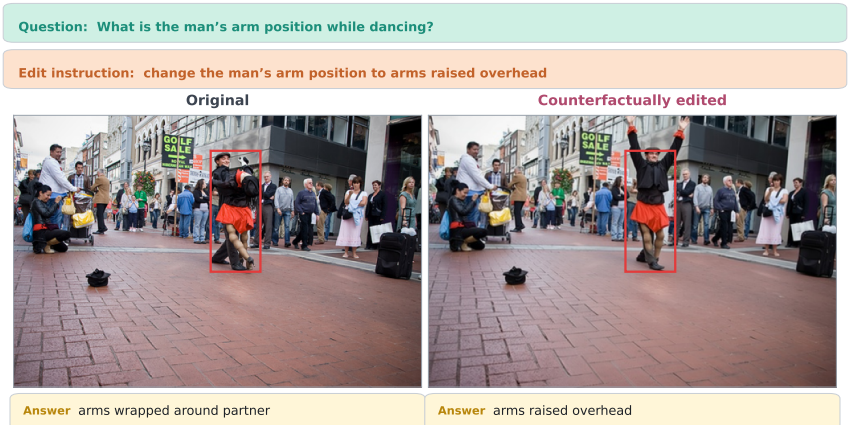


Figure 6: Changing what the subject is doing.

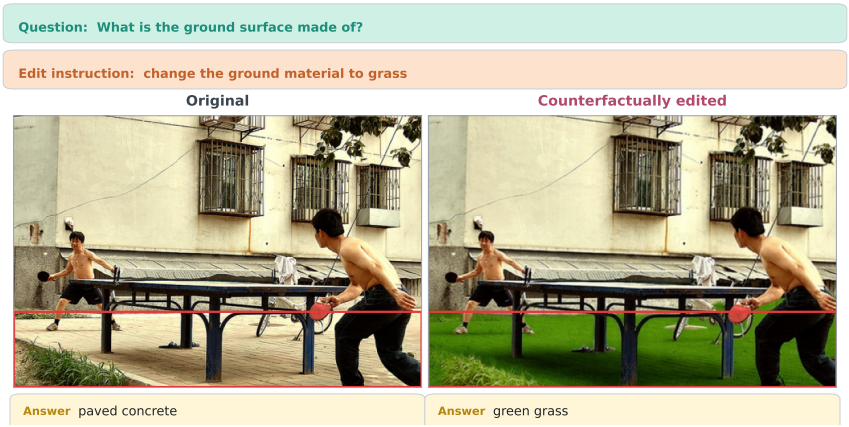


Figure 7: Swapping the material of a surface.

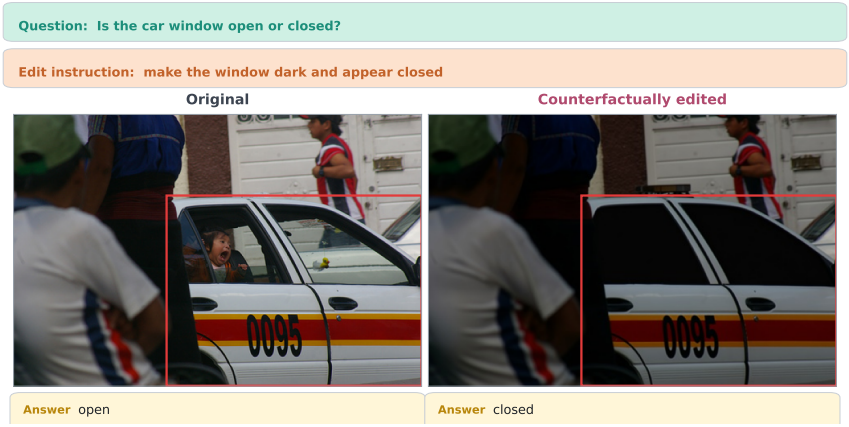


Figure 8: Switching an object between two states.

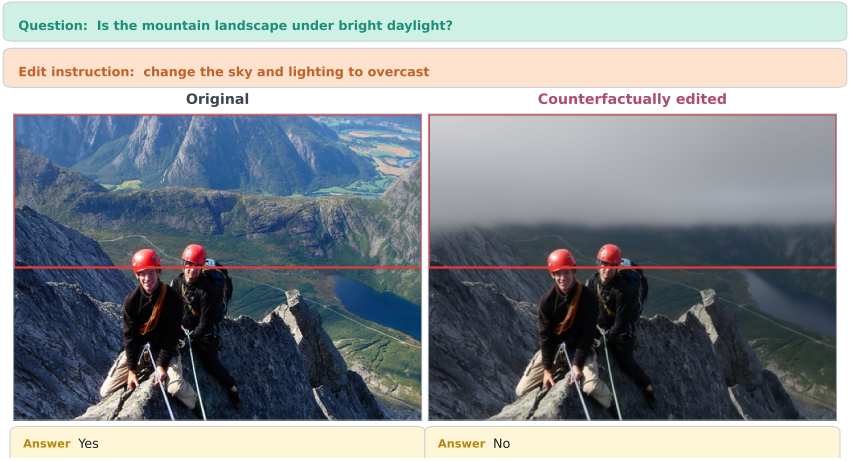


Figure 9: Changing weather or time of day.



Figure 10: Replacing the pattern on a surface.