UNDERSTAND, IMAGINE, THEN REASON: THINKING WITH FORESIGHT IN UNIFIED LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have made substantial progress in visual understanding, yet reasoning about spatial relationships, object interactions, and state changes remains challenging. Thinking about images primarily operates in text, thinking with images often relies on external tools for visual evidence, and existing latent visual reasoning methods largely focus on static representations. Unified multimodal models offer an opportunity to simulate scene dynamics, yet how to leverage these dynamics for visual reasoning remains underexplored. We propose **Understand-Imagine-Reason (UnI-R)**, a framework that feeds internally simulated future dynamics back into reasoning, forming an understanding, imagination, and reasoning loop within a unified latent space. We first introduce Anchor-Conditioned Dynamic Coverage to select complementary future dynamics beyond the observed static scene, while a Foresight-to-Reasoning Bridge aggregates and integrates them into the reasoning stream. During training, Consistent and Discriminative Foresight Learning promotes consistency across rollouts and discrimination across inputs. Contrastive Future-Grounded Self-Distillation further aligns predictions based on imagined futures with teacher predictions grounded in actual future observations while contrasting them with those induced by counterfactual futures. UnI-R empirically improves visual reasoning across both dynamic-centric and general scenarios, achieving an average gain of 8.76% on BabyVision and an 8.34% improvement on MME-RealWorld, demonstrating the promise of leveraging dynamics for reasoning in unified latent space.

1 INTRODUCTION

Large Language Models (LLMs) (Bai et al., 2023a; Chowdhery et al., 2023; Dubey et al., 2024; Laurençon et al., 2023) have demonstrated strong capabilities in language understanding, generation, and reasoning, motivating efforts to extend these capabilities beyond text. By integrating vision encoders with LLMs, Multimodal Large Language Models (MLLMs) (Bai et al., 2023b; Chen et al., 2023; Achiam et al., 2023; Li et al., 2023; Liu et al., 2024a; Chen et al., 2024b) have made significant progress in visual understanding and multimodal reasoning. However, recognizing visual content does not necessarily entail understanding spatial relations, object interactions, and state dynamics, leaving complex visual reasoning a persistent challenge for MLLMs.

Existing research has explored several ways to enhance visual reasoning. As shown in Figure 1, *Thinking about image* (Zhang et al., 2023; Huang et al., 2026; Sarch et al., 2026) mainly reasons in language, where translating visual relations into text may lose geometric and motion cues (Gupta & Kembhavi, 2023; Su et al., 2025). *Thinking with image* (Izadi et al., 2025; Gao et al., 2025; Li et al., 2025; Wu et al., 2024a; Li et al., 2026b) introduces cropping, annotation, or intermediate image generation, but often depends on external tools and discrete, rule-based visual manipulations (Su et al., 2025; Li et al., 2026a; Wang et al., 2026b). *Latent visual reasoning* (Li et al., 2026a; Wang et al., 2026b; Yang et al., 2026; Qin et al., 2025) instead moves intermediate steps into continuous latent spaces, reducing reliance on explicit intermediates. However, existing methods largely focus on local reconstruction or static semantic alignment, with supervision often derived from auxiliary images or crops (Schulze Buschoff et al., 2025; Park et al., 2020). Such alignment does not ensure that the latent representations capture meaningful scene dynamics or effectively support subsequent reasoning. These limitations motivate moving beyond static visual evidence toward representing and leveraging scene dynamics to support reasoning within a unified and interpretable latent space.

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

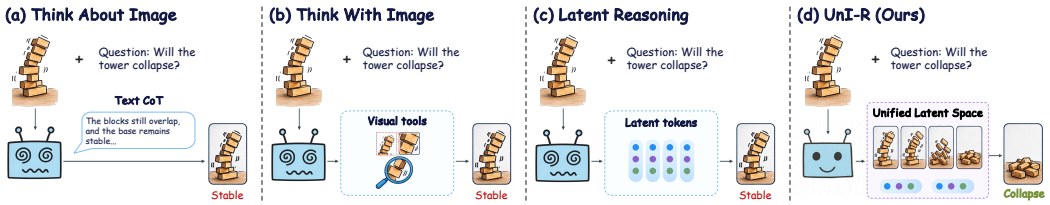


Figure 1: **Illustrative comparison of visual reasoning paradigms**, highlighting how UnI-R leverages imagined future dynamics within a unified latent space to reason beyond static observations.

Human visual reasoning offers a useful perspective on this challenge. When observing an unstable stack of objects, people can imagine how it might collapse to assess its stability (Park et al., 2020; Chandra et al., 2023; Bill et al., 2022; de Vries & Wurm, 2023; Balaban & Ullman, 2025) . Even for static visual questions, they may mentally zoom in on relevant regions to focus on details that inform the answer Battaglia et al. (2013); Fischer et al. (2016); Kourtzi & Kanwisher (2000). Such reasoning brings observations and prior knowledge together within a coherent internal representation to explore relevant visual changes, rather than arbitrarily invent unseen details. This suggests that dynamic visual reasoning may support not only future prediction, but also questions about static images that benefit from closer inspection or implicit visual transformations.

Advances in unified modeling of multimodal understanding and generation provide a new foundation for bringing this idea to MLLMs (Agarwal et al., 2026; Fu et al., 2026; Deng et al., 2025; Xie et al., 2026). By combining multimodal understanding with video generation in a single architecture, these models can interpret observed scenes and simulate how they might evolve. Beyond enabling both question answering and visual generation, this integration offers a common representational foundation for connecting the two capabilities (Agarwal et al., 2026; Fu et al., 2026). However, their coexistence does not automatically allow imagined dynamics to inform reasoning. How to employ such dynamics within a unified latent space to support visual reasoning remains underexplored.

We propose **Understand-Imagine-Reason (UnI-R)**, a training framework that connects future dynamics and visual reasoning within a unified latent space. As shown in Figure 1, UnI-R feeds internally imagined future dynamics back into the reasoning process, forming a loop of understanding, imagination, and further reasoning. To extract useful foresight from the generation branch, *Anchor-Conditioned Dynamic Coverage* (ACDC) selects informative and complementary latent dynamics from imagined futures, which are then integrated into the reasoning stream through a *Foresight-to-Reasoning Bridge* (F2R). We first train the model with *Consistent and Discriminative Foresight Learning* (CDFL), enabling it to learn stable and discriminative foresight representations for downstream reasoning through supervised fine-turning. We further introduce *Contrastive Future-Grounded Self-Distillation* (CFGSD), which improves the reliability of reasoning with imagined futures through guidance from actual and counterfactual future observations. These additional future observations are used only during training, while inference relies solely on the model's own imagined futures. Together, these components allow visual generation to serve as an internal source of dynamic evidence for visual reasoning. Our contributions are summarized as follows:

- We introduce **UnI-R**, a framework that connects imagined future dynamics with visual reasoning in a unified latent space. UnI-R achieves an average improvement of **8.76%** over the baseline on BabyVision (Chen et al., 2026) and a gain of **8.34%** on MME-RealWorld (Zhang et al., 2025), demonstrating that internally imagined dynamics can benefit both dynamic reasoning and general visual understanding across diverse challenging visual reasoning benchmarks.
- We propose ACDC, which selects complementary latent dynamics from future trajectories relative to the current observation, providing compact yet informative dynamic cues for reasoning.
- We develop F2R, which aggregates the selected future dynamics and injects them into the reasoning stream, enabling imagined scene evolution to directly inform subsequent predictions.
- We introduce a two-stage training strategy. CDFL learns consistent and discriminative foresight representations through SFT, while CFGSD further improves reasoning with imagined futures through contrastive guidance from actual and counterfactual future observations.

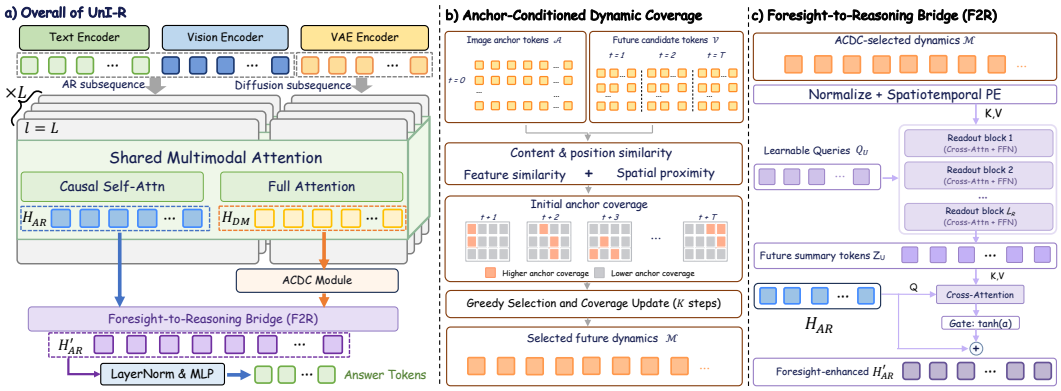


Figure 2: **Overview of UnI-R.** (a) UnI-R connects understanding, imagination, and reasoning within a unified latent space, feeding internally generated future dynamics back into the reasoning stream. (b) ACDC select complementary future states beyond the coverage provided by the observed frame. (c) F2R Bridge uses learnable queries to aggregate the selected states into future summary tokens .

2 Related Work

2.1 Multimodal Large Language Models

Multimodal Large Language Models extend language modeling to visual inputs, with early work (Dai et al., 2023; Liu et al., 2024a) connecting visual encoders to LLMs through visual instruction tuning. Subsequent models Liu et al. (2024b); Li et al. (2024); Chen et al. (2024b); Wu et al. (2024b); Bai et al. (2025); Achiam et al. (2023) further improve visual perception, video understanding, and multimodal reasoning. Building on this progress, unified multimodal models integrate understanding and generation within a single architecture. BAGEL (Deng et al., 2025) unifies multimodal understanding and image generation, Show-o2 (Xie et al., 2026) extends this paradigm to image and video generation, and Lance (Fu et al., 2026) jointly supports image and video understanding, generation, and editing. Cosmos 3 (Agarwal et al., 2026) further couples multimodal reasoning with diffusion-based scene simulation. However, architectural unification alone does not ensure that generated dynamics contribute to subsequent reasoning, motivating the use of imagined futures for visual reasoning within a unified latent space.

2.2 Visual Reasoning in MLLMs

Research on visual reasoning in MLLMs has explored different forms of intermediate computation. *Thinking about image* (Zhang et al., 2023; Huang et al., 2026; Sarch et al., 2026; He et al., 2025) reasons in language, but verbalizing visual relations can lose details needed for precise judgments. *Thinking with image* (Izadi et al., 2025; Gao et al., 2025; Li et al., 2025; Wu et al., 2024a; Li et al., 2026b; Zhang et al., 2026; Zhong et al., 2026) introduces additional visual evidence through cropping, annotation, or image generation, but often relies on external tools or explicit visual intermediates. *Latent visual reasoning* (Li et al., 2026a; Wang et al., 2026b; Yang et al., 2026; Qin et al., 2025) moves intermediate reasoning into continuous embeddings, reducing reliance on explicit representations. However, existing methods mainly focus on reconstructing or distilling static visual information rather than modeling scene evolution, and alignment with prescribed targets does not necessarily ensure that learned representations effectively support subsequent reasoning. These limitations motivate UnI-R, which connects world generation and visual reasoning in a unified latent space so that imagined future dynamics can inform reasoning beyond the observed static scene.

3 Method

UnI-R forms a loop of understanding, imagination, and reasoning within a unified latent space. Its autoregressive reasoner and diffusion generator interact through shared multimodal attention,

$$H_{\mathrm{AR}} = \mathrm{Attn}_{\mathrm{causal}}(Q_{\mathrm{AR}}, K_{\mathrm{AR}}, V_{\mathrm{AR}}), \quad H_{\mathrm{DM}} = \mathrm{Attn}_{\mathrm{full}}(Q_{\mathrm{DM}}, [K_{\mathrm{AR}}, K_{\mathrm{DM}}], [V_{\mathrm{AR}}, V_{\mathrm{DM}}]). \quad (1)$$

This asymmetric structure allows generation to depend on the reasoner, but not generated dynamics to directly inform reasoning. As shown in Figure 2(a), UnI-R closes this loop by feeding imagined future dynamics back into the reasoning stream. Given an image and question, ACDC selects complementary latent states from simulated futures, while the F2R Bridge aggregates and integrates them into H_{AR} for subsequent reasoning and answer prediction. UnI-R is trained through CDFL followed by CFGSD, with actual future observations used only as training-time guidance.

3.1 Anchor-Conditioned Dynamic Coverage

Figure 2(b) shows that ACDC selects complementary future states according to the coverage they add beyond the observed static scene. Let $\mathcal{A}$ denote the image anchor tokens and $\mathcal{V}$ the candidates from T imagined future frames in H_{DM}. Each token i contains a hidden state h_i and its position (t_i, u_i, v_i). The entire budget $K \leq |\mathcal{V}|$ is allocated to future tokens, while anchors only as reference.

We compare local states using normalized features $z_i = \mathrm{Normalize}(h_i)$ and normalized grid cell centers $p_i = ((u_i + 0.5)/W, (v_i + 0.5)/H)$ on an $H \times W$ grid. Their similarity is

$$k(i,j) = \exp\big(z_i^\top z_j - 1 - \|p_i - p_j\|_2^2\big). \quad (2)$$

Normalization uses the L_2 norm. The anchor coverage of candidate i is $b_i = \max_{a\in\mathcal{A}} k(i,a)$. For a selected set $\mathcal{S} \subseteq \mathcal{V}$, we maximize

$$F(\mathcal{S} \mid \mathcal{A}) = \sum_{i\in\mathcal{V}} \left[\max_{j\in\mathcal{A}\cup\mathcal{S}} k(i,j) - b_i\right] \quad (3)$$

subject to $|\mathcal{S}| \leq K$. This objective rewards complementary coverage of future changes rather than difference. Algorithm 1 greedily selects the candidate with the largest marginal gain and updates the coverage. Selection operates across all future frames without fixed frame quotas. The output is $\mathcal{M} = \{(h_i, t_i, u_i, v_i) \mid i \in \mathcal{S}_K\}$. The detailed derivation is provided in Appendix B.

Algorithm 1 ACDC states selection

Require: Anchor tokens $\mathcal{A}$, future tokens $\mathcal{V}$, grid $H \times W$, budget $K \leq |\mathcal{V}|$
Ensure: Selected future dynamic memory $\mathcal{M}$
Compute z_i and p_i for $i \in \mathcal{A} \cup \mathcal{V}$
Define $k(i,j)$ by Eq. equation 2
for all $i \in \mathcal{V}$ **do**
  $b_i \leftarrow \max_{a\in\mathcal{A}} k(i,a)$, $c_i^{(0)} \leftarrow b_i$
end for
$\mathcal{S}_0 \leftarrow \varnothing$
for $r = 0, \ldots, K-1$ **do**
  For each $j \in \mathcal{V} \setminus \mathcal{S}_r$, compute $\Delta(j \mid \mathcal{S}_r) \leftarrow \sum_{i\in\mathcal{V}} \max\left(0, k(i,j) - c_i^{(r)}\right)$
  Select $j_r \in \mathrm{argmax}_{j\in\mathcal{V}\setminus\mathcal{S}_r} \Delta(j \mid \mathcal{S}_r)$
  $\mathcal{S}_{r+1} \leftarrow \mathcal{S}_r \cup \{j_r\}$
  for all $i \in \mathcal{V}$ **do**
    $c_i^{(r+1)} \leftarrow \max(c_i^{(r)}, k(i, j_r))$
  end for
end for
$\mathcal{M} \leftarrow \{(h_i, t_i, u_i, v_i) \mid i \in \mathcal{S}_K\}$
return $\mathcal{M}$

3.2 Foresight-to-Reasoning Bridge

The Foresight-to-Reasoning Bridge (F2R) summarizes the selected future dynamic memory $\mathcal{M}$ for autoregressive (AR) reasoning. For each $i \in \mathcal{S}_K$, we project the original hidden state h_i and add its spatiotemporal position encoding,

$$m_i = \mathrm{Linear}(h_i) + \mathrm{PE}(t_i, u_i, v_i), \quad M = \mathrm{Stack}_{i\in\mathcal{S}_K}(m_i) \in \mathbb{R}^{K\times d_{\mathrm{AR}}}, \quad (4)$$

where d_{AR} is the AR hidden dimension. We initialize R learnable queries as $Z^{(0)} = Q_U \in \mathbb{R}^{R\times d_{\mathrm{AR}}}$ and update them through L_R readout blocks,

$$\widetilde{Z}^{(\ell)} = \mathrm{LN}\left(Z^{(\ell-1)} + \mathrm{MHA}\left(Z^{(\ell-1)}, M, M\right)\right), \quad Z^{(\ell)} = \mathrm{LN}\left(\widetilde{Z}^{(\ell)} + \mathrm{FFN}\left(\widetilde{Z}^{(\ell)}\right)\right), \quad (5)$$

where $\ell = 1, \ldots, L_R$. All queries attend to the complete memory without predefined frame or object assignments, yielding the future summaries $Z_U = Z^{(L_R)}$.

Given L AR states $H_{\mathrm{AR}} \in \mathbb{R}^{L\times d_{\mathrm{AR}}}$, we integrate these summaries through gated cross attention,

$$\Delta H_{\mathrm{AR}} = \mathrm{MHA}\left(H_{\mathrm{AR}}, Z_U, Z_U\right), \quad H'_{\mathrm{AR}} = H_{\mathrm{AR}} + \tanh(\alpha)\,\Delta H_{\mathrm{AR}}, \quad (6)$$

where α is a learnable scalar shared across tokens and channels. The bridge is inserted after the attention sublayer of the final Transformer block and before its original LayerNorm and MLP. The augmented states then continue through the backbone for reasoning and answer generation.

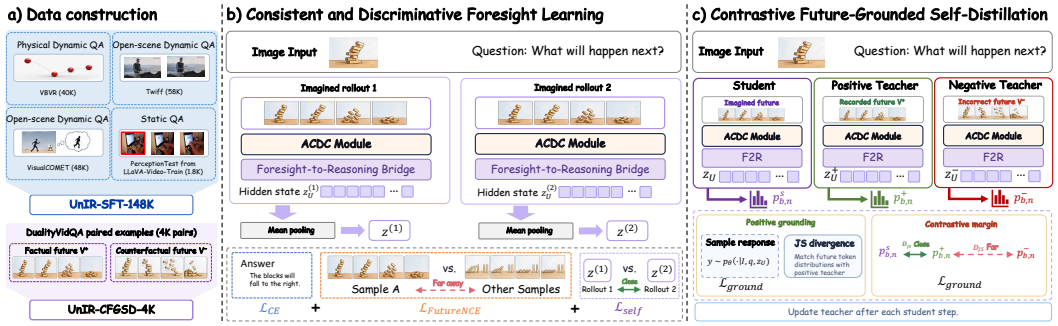


Figure 3: **Data construction and training pipeline of UnI-R.** (a) UnIR-SFT-148K combines dynamic and static VQA, while UnIR-CFGSD-4K contains factual and counterfactual pairs. (b) CDFL combines response supervision, FutureNCE, and self alignment to learn consistent and discriminative future representations. (c) CFGSD aligns student predictions with those of the factual future teacher while encouraging separation from those of the counterfactual future teacher.

3.3 SFT Stage: Consistent and Discriminative Foresight Learning

We first apply Consistent and Discriminative Foresight Learning (CDFL) to teach UnI-R to use imagined dynamics for visual reasoning. We construct UnIR-SFT-148K for foresight reasoning, covering dynamic scene evolution and static visual understanding. For dynamic VQA, we combine 40K physical interaction examples from VBVR (Wang et al., 2026a), with 58K dynamic scene examples from Twiff (Liu et al., 2026) and 48K open world examples from VisualCOMET (Park et al., 2020). For static VQA, we use the 1.8K PerceptionTest subset of LLaVA-Video-Train Zhang et al. (2024) to construct question and answer pairs with the first frame as the visual input.

Different random seeds initialize the diffusion process with distinct noise, yielding different imagined future rollouts conditioned on the same image and question (I, q). ACDC and F2R process into future summaries $Z_U^{(1)}$ and $Z_U^{(2)}$. Training combines response supervision with FutureNCE and Self-Align. Given a target assistant response $y = (y_1, \ldots, y_{N_y})$, we minimize

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{2B}\sum_{b=1}^{B}\frac{1}{N_{y_b}}\sum_{s=1}^{2}\sum_{n=1}^{N_{y_b}}\log p_\theta\Big(y_{b,n} \mid I_b, q_b, y_{b,<n}, Z_{U,b}^{(s)}\Big). \tag{7}$$

where the expectation is over training examples and sampled rollouts. Only assistant tokens, including any retained explanation and final answer, contribute to the loss. For a distributed batch of B examples, we mean-pool each future summary and apply L_2 normalization,

$$\bar{z}_b^{(s)} = \frac{1}{R}\sum_{r=1}^{R}[Z_{U,b}^{(s)}]_r \in \mathbb{R}^{d_{\mathrm{AR}}}, \quad e_b^{(s)} = \mathrm{Normalize}\Big(\bar{z}_b^{(s)}\Big), \quad s \in \{1,2\}. \tag{8}$$

FutureNCE aligns embeddings from two rollouts of the same input while separating those from different inputs. For each example b, the negative set $\mathcal{N}_b$ contains embeddings from both rollouts of all other examples in the distributed batch. We use the symmetric objective

$$\mathcal{L}_{\mathrm{FutureNCE}} = -\frac{1}{2B}\sum_{b=1}^{B}\sum_{s=1}^{2}\log\frac{\exp\Big((e_b^{(s)})^\top e_b^{(3-s)}\Big)}{\exp\Big((e_b^{(s)})^\top e_b^{(3-s)}\Big) + \sum_{e^-\in\mathcal{N}_b}\exp\Big((e_b^{(s)})^\top e^-\Big)}, \tag{9}$$

Self-Align directly matches the unnormalized pooled summaries without negatives,

$$\mathcal{L}_{\mathrm{self}} = \frac{1}{Bd_{\mathrm{AR}}}\sum_{b=1}^{B}\Big\|\bar{z}_b^{(1)} - \bar{z}_b^{(2)}\Big\|_2^2. \tag{10}$$

These auxiliary objectives require neither token-level correspondence between ACDC selections nor ground-truth future observations. The overall objective is

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{FutureNCE}} + \mathcal{L}_{\mathrm{self}}. \tag{11}$$

We optimize F2R and MLP in the final Transformer block, keeping the rest of the backbone frozen.

3.4 CONTRASTIVE FUTURE-GROUNDED SELF-DISTILLATION

Standard On-Policy Self-Distillation (OPSD) (Zhao et al., 2026) lacks explicit negative-teacher supervision, yet rejecting misleading future dynamics is important for reliable reasoning. We therefore introduce Contrastive Future-Grounded Self-Distillation (CFGSD) to ground predictions in factual dynamics while contrasting them with those induced by counterfactual futures. We construct UnIR-CFGSD-4K using 4K pairs of factual and counterfactual examples from DualityVidQA (Huang et al., 2025). These pairs share similar scene structures but differ in causal dynamics.

For each image and question (I, q), let V^+ and V^- denote the factual and counterfactual futures. The student uses its own imagined future, while the positive and negative teachers receive V^+ and V^-. Both teachers share an exponential moving average (EMA) copy $\bar{\theta}$ of the student parameters θ, initialized from the CDFL checkpoint, and differ only in their future conditions. Teacher futures are VAE-encoded and processed through the diffusion state extraction interface. All branches apply ACDC with budget K, followed by F2R, yielding student summaries $\widehat{Z}_U$ and teacher summaries Z_U^+ and Z_U^-. The student samples a response $Y_b \sim p_\theta(\cdot \mid I_b, q_b, \widehat{Z}_{U,b})$. At each position n, all three branches predict over the full vocabulary given the same student prefix $Y_{b,<n}$,

$$p^S_{b,n} = p_\theta\Big(\cdot \mid I_b, q_b, Y_{b,<n}, \widehat{Z}_{U,b}\Big), \quad p^{\pm}_{b,n} = \mathrm{sg}\Big[p_{\bar{\theta}}\Big(\cdot \mid I_b, q_b, Y_{b,<n}, Z^{\pm}_{U,b}\Big)\Big], \tag{12}$$

where sg stops gradients, and the sampled response is treated as fixed context. We define $d^{\pm}_{b,n} = D_{\mathrm{JS}}(p^S_{b,n}, p^{\pm}_{b,n})$ and $d^{+-}_{b,n} = D_{\mathrm{JS}}(p^+_{b,n}, p^-_{b,n})$ using Jensen–Shannon divergence.

We filter supervision using separately generated deterministic teacher responses. Let $\mathcal{I}_+$ contain all student response positions, including reasoning and final answer tokens, from examples whose positive teacher answers correctly. The subset $\mathcal{I}_{\mathrm{ctr}} \subseteq \mathcal{I}_+$ additionally requires a verified incorrect future V_b^- and a valid incorrect negative-teacher answer. The grounding and contrastive losses are

$$\mathcal{L}_{\mathrm{ground}} = \frac{\sum_{(b,n)\in\mathcal{I}_+} d^+_{b,n}}{\max(1, |\mathcal{I}_+|)}, \quad \mathcal{L}_{\mathrm{ctr}} = \frac{\sum_{(b,n)\in\mathcal{I}_{\mathrm{ctr}}} \Big[d^+_{b,n} - d^-_{b,n} + d^{+-}_{b,n}\Big]_+}{\max(1, |\mathcal{I}_{\mathrm{ctr}}|)}, \tag{13}$$

where $[x]_+ = \max(0, x)$. The contrastive penalty vanishes when the teacher distributions coincide or the student matches the positive teacher. Examples without valid negative supervision contribute only to grounding when positive supervision is valid. The overall objective is $\mathcal{L}_{\mathrm{CFGSD}} = \mathcal{L}_{\mathrm{ground}} + \mathcal{L}_{\mathrm{ctr}}$. We optimize F2R and the MLP in the final Transformer block, keeping the rest of the backbone frozen, and update the shared teacher by EMA after each student update.

4 EXPERIMENT

Baselines. We compare UnI-R with representative methods from *Thinking About Images*, *Thinking With Images*, and *Latent Visual Reasoning*. The first category includes Vision-R1 (Huang et al., 2026) and ViGoRL (Sarch et al., 2026), while the second includes VISER (Izadi et al., 2025), ICoT (Gao et al., 2025), ControlMLLM (Wu et al., 2024a), DyFo (Li et al., 2025), CoFFT (Zhang et al., 2026), DeepScan (Li et al., 2026b), and FOCUS (Zhong et al., 2026). For latent visual reasoning, we compare against CoVT (Qin et al., 2025), LVR (Li et al., 2026a), and Monet (Wang et al., 2026b). For a fair comparison, we implement UnI-R and all baselines on Cosmos3-Nano (Agarwal et al., 2026), which is built upon Qwen3-VL-8B (Bai et al., 2025), a dense Transformer with 8B parameters. More details can be found in Appendix C.

Benchmarks. We evaluate UnI-R on two task categories. *Dynamic-Centric Image Understanding Tasks* include VTBench (Tong et al., 2026), Visual Cognitive Tasks (Schulze Buschoff et al., 2025), BabyVision (Chen et al., 2026), OpenWorldQA(Zhou et al., 2026), and CausalSpatial(Ma et al., 2026). *General Image Understanding Tasks* include MMBench (Liu et al., 2024c), MMStar (Chen et al., 2024a), MMVet (Yu et al., 2024), V* (Wu & Xie, 2024), HRBench4K/8K (Wang et al., 2025), MME-RealWorld (Zhang et al., 2025), RealworldQA (xAI, 2024), SpatialEval (Wang et al., 2024).

Hyperparameters. All models are trained and evaluated on $8\times$ A100 GPUs. We perform CDFL for two epochs, followed by one epoch of CFGSD training for UnI-R. ACDC selects a total of $K = 256$ hidden states across $T = 32$ generated futures, and the F2R Bridge uses $R = 32$ learnable queries and $L_R = 3$ readout blocks. These configurations remain the same during training and inference. We use a learning rate of 1×10^{-5} and a weight decay of 0.01. Details are in Appendix C.3.

Table 1: **Performance comparison on Dynamic-Centric Image Understanding Tasks.** Blue denotes improvement, gray denotes degradation, the best-performing results are highlighted in bold.

Models	BabyVision					Visual cognitive tasks		VTBench	OpenWorldQA	CausalSpatial
	Fine-grained Discrimination	Visual Tracking	Spatial Perception	Visual Pattern Recognition	Overall	Block_towers	Michotte			
Qwen3-VL-8B	7.36	7.23	6.59	9.80	7.47	51.48	19.81	23.22	32.00	36.96
Cosmos3	7.98 (0.00)	8.43 (0.00)	9.89 (0.00)	7.84 (0.00)	8.51 (0.00)	53.06 (0.00)	19.02 (0.00)	23.87 (0.00)	35.80 (0.00)	38.44 (0.00)
Think About Image										
+Vision-R1	9.20 (+1.22)	8.43 (0.00)	10.99 (+1.10)	11.76 (+3.92)	9.79 (+1.28)	52.30 (-0.76)	19.45 (+0.43)	23.87 (0.00)	36.00 (+0.20)	38.04 (-0.40)
+ViGoRL	9.82 (+1.84)	7.23 (-1.20)	8.79 (-1.10)	11.76 (+3.92)	9.28 (+0.77)	53.97 (+0.91)	18.38 (-0.64)	22.90 (-0.97)	35.20 (-0.60)	37.55 (-0.89)
Think With Image										
+ControlMLLM	8.59 (+0.61)	10.84 (+2.41)	9.89 (0.00)	13.73 (+5.89)	10.05 (+1.54)	51.27 (-1.79)	18.90 (-0.12)	23.55 (-0.32)	36.20 (+0.40)	38.93 (+0.49)
+DyFo	9.82 (+1.84)	8.43 (0.00)	10.99 (+1.10)	9.80 (+1.96)	9.79 (+1.28)	52.96 (-0.10)	19.93 (+0.91)	24.52 (+0.65)	37.40 (+1.60)	38.83 (+0.39)
+CoFFT	7.98 (0.00)	9.64 (+1.21)	13.19 (+3.30)	5.88 (-1.96)	9.28 (+0.77)	53.61 (+0.55)	20.85 (+1.83)	25.16 (+1.29)	36.00 (+0.20)	40.02 (+1.58)
+DeepScan	7.36 (-0.62)	9.64 (+1.21)	7.69 (-2.20)	13.73 (+5.89)	8.76 (+0.25)	53.75 (+0.69)	20.56 (+1.54)	24.19 (+0.32)	36.20 (+0.40)	37.85 (-0.59)
+FOCUS	10.43 (+2.45)	10.84 (+2.41)	8.79 (-1.10)	7.84 (0.00)	9.79 (+1.28)	54.60 (+1.54)	21.87 (+2.85)	25.48 (+1.61)	36.40 (+0.60)	38.44 (0.00)
+VISER	9.20 (+1.22)	7.23 (-1.20)	10.99 (+1.10)	11.76 (+3.92)	9.54 (+1.03)	52.57 (-0.49)	21.02 (+2.00)	26.77 (+2.90)	37.60 (+1.80)	40.32 (+1.88)
+ICoT	8.59 (+0.61)	8.43 (0.00)	9.89 (0.00)	9.80 (+1.96)	9.02 (+0.51)	53.21 (+0.15)	20.34 (+1.32)	25.16 (+1.29)	37.00 (+1.20)	39.53 (+1.09)
Latent Reasoning Method										
+CoVT	10.43 (+2.45)	12.05 (+3.62)	12.09 (+2.20)	7.84 (0.00)	10.82 (+2.31)	55.03 (+1.97)	21.31 (+2.29)	24.52 (+0.65)	36.40 (+0.60)	41.01 (+2.57)
+LVR	9.20 (+1.22)	10.84 (+2.41)	8.79 (-1.10)	11.76 (+3.92)	9.79 (+1.28)	55.66 (+2.60)	21.29 (+2.27)	27.10 (+3.23)	37.00 (+1.20)	39.62 (+1.18)
+Monet	11.04 (+3.06)	12.05 (+3.62)	9.89 (0.00)	9.80 (+1.96)	10.82 (+2.31)	56.19 (+3.13)	22.18 (+3.16)	25.81 (+1.94)	38.20 (+2.40)	41.11 (+2.67)
Our Method										
+UnI-R	**17.18** (+9.20)	**16.87** (+8.44)	**18.68** (+8.79)	**15.69** (+7.85)	**17.27** (+8.76)	**62.35** (+9.29)	**26.53** (+7.51)	**31.29** (+7.42)	**41.40** (+5.60)	**43.28** (+4.84)

Table 2: **Performance comparison on General Image Understanding Tasks.**

Models	Comprehensive				RealWorld				Spatial
	MMBench	MMStar	MMVet	V*	HRBench4K	HRBench8K	MME-RealWorld	RealworldQA	SpatialEval
Qwen3-VL-8B	82.96	63.65	73.10	73.82	74.13	68.50	44.71	70.31	64.05
Cosmos3-Nano	82.73 (0.00)	62.57 (0.00)	70.08 (0.00)	75.92 (0.00)	75.38 (0.00)	70.13 (0.00)	44.40 (0.00)	71.45 (0.00)	64.69 (0.00)
Think About Image									
+Vision-R1	83.09 (+0.36)	63.26 (+0.69)	77.45 (+7.37)	72.77 (-3.15)	74.38 (-1.00)	69.38 (-0.75)	43.46 (-0.94)	70.48 (-0.97)	65.14 (+0.45)
+ViGoRL	82.81 (+0.08)	62.94 (+0.37)	**78.90** (+8.82)	73.82 (-2.10)	74.63 (-0.75)	69.25 (-0.88)	44.09 (-0.31)	69.73 (-1.72)	64.80 (+0.11)
Think With Image									
+ControlMLLM	83.18 (+0.45)	64.37 (+1.80)	74.32 (+4.24)	75.39 (-0.53)	74.75 (-0.63)	69.00 (-1.13)	43.78 (-0.62)	70.95 (-0.50)	63.66 (-1.03)
+DyFo	83.99 (+1.26)	64.53 (+1.96)	73.75 (+3.67)	75.92 (0.00)	75.13 (-0.25)	69.88 (-0.25)	45.37 (+0.97)	70.18 (-1.27)	64.79 (+0.10)
+CoFFT	83.59 (+0.86)	65.70 (+3.13)	74.84 (+4.76)	75.39 (-0.53)	75.00 (-0.38)	68.63 (-1.50)	45.85 (+1.45)	71.90 (+0.45)	64.51 (-0.18)
+DeepScan	83.27 (+0.54)	65.01 (+2.44)	74.36 (+4.28)	74.34 (-1.58)	76.38 (+1.00)	68.25 (-1.88)	46.88 (+2.48)	72.01 (+0.56)	64.34 (-0.35)
+FOCUS	84.32 (+1.59)	65.25 (+2.68)	75.63 (+5.55)	76.96 (+1.04)	75.88 (+0.50)	69.13 (-1.00)	48.25 (+3.85)	71.37 (-0.08)	65.21 (+0.52)
+VISER	84.56 (+1.83)	64.68 (+2.11)	75.30 (+5.22)	75.92 (0.00)	76.75 (+1.37)	70.25 (+0.12)	46.78 (+2.38)	71.58 (+0.13)	65.74 (+1.05)
+ICoT	83.05 (+0.32)	65.36 (+2.79)	74.80 (+4.72)	76.44 (+0.52)	76.63 (+1.25)	70.63 (+0.50)	46.30 (+1.90)	71.96 (+0.51)	65.48 (+0.79)
Latent Reasoning Method									
+CoVT	84.22 (+1.49)	64.03 (+1.46)	75.19 (+5.11)	78.01 (+2.09)	77.00 (+1.62)	70.50 (+0.37)	48.09 (+3.69)	71.03 (-0.42)	64.61 (-0.08)
+LVR	85.16 (+2.43)	64.89 (+2.32)	74.86 (+4.78)	76.96 (+1.04)	77.25 (+1.87)	72.13 (+2.00)	47.94 (+3.54)	72.46 (+1.01)	64.97 (+0.28)
+Monet	85.34 (+2.61)	65.28 (+2.71)	75.27 (+5.19)	78.53 (+2.61)	77.63 (+2.25)	70.25 (+0.12)	49.04 (+4.64)	72.35 (+0.90)	65.16 (+0.47)
Our Method									
+UnI-R	**87.08** (+4.35)	**67.13** (+4.56)	78.66 (+8.58)	**81.16** (+5.24)	**80.88** (+5.50)	**74.13** (+4.00)	**52.74** (+8.34)	**74.90** (+3.45)	**68.54** (+3.85)

4.1 Main Result

Dynamic-centric Image Reasoning Benchmarks. As shown in Table 1, UnI-R outperforms all compared methods across every reported benchmark and all four BabyVision ability categories. On BabyVision, it improves the overall score by 8.76 points and surpasses the strongest competing result by 6.45 points, demonstrating a substantial advantage over existing reasoning methods. The gains extend beyond BabyVision, with notable improvements on Block_towers and VTBench, alongside leading results on OpenWorldQA, and CausalSpatial. These consistent improvements support the effectiveness of integrating imagined dynamics into visual reasoning within a unified latent space.

General Image Understanding Tasks. Table 2 shows that UnI-R extends its advantages beyond dynamic tasks, improving over Cosmos3-Nano on all nine benchmarks and achieving the best results on eight. On comprehensive benchmarks, UnI-R raises the V* score from 75.92 to 81.16. The benefits are particularly pronounced on MME-RealWorld, where performance increases from 44.40 to 52.74. The advantage also extends to spatial reasoning, with SpatialEval improving from 64.69 to 68.54. These results demonstrate that imagined dynamics provide broadly useful information for visual reasoning, benefiting not only dynamic-centric tasks but also general, real-world, and spatial image understanding. More detailed experimental results are provided in Appendix D.

Table 3: Component ablation of UnI-R.

Model	BabyVision	VTBench	V*	MME-RW	SpatialEval
Baseline					
Cosmos3	8.51	23.87	75.92	44.40	64.69
+ Standard SFT	8.76	24.84	74.87	45.18	64.85
Our Method					
UnI-R	**17.27**	**31.29**	**81.16**	**52.74**	**68.54**
w/o ACDC	14.18	28.71	79.58	50.50	67.80
w/o CFGSD	15.46	29.68	79.05	50.81	68.16

Table 4: Loss function ablation of UnI-R.

Model	BabyVision	VTBench	V*	MME-RW	SpatialEval
UnI-R	**17.27**	**31.29**	**81.16**	**52.74**	**68.54**
CDFL Stage					
w/o $\mathcal{L}_{\text{FutureNCE}}$	15.72	28.06	77.49	47.94	65.70
w/o $\mathcal{L}_{\text{self-align}}$	16.24	28.39	78.53	49.45	68.11
CFGSD Stage					
w/o Contrastive	16.49	29.03	78.53	50.03	67.95

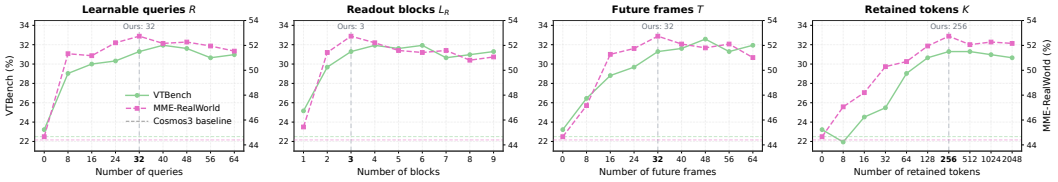


Figure 4: Ablation on the hyperparameters of UnI-R.

4.2 Ablation and Analysis

Ablation of the Main Components. As shown in Table 3, standard SFT on the same training data (UnIR-SFT-148K) as UnI-R yields only modest gains on BabyVision and VTBench. This indicates that the training data alone cannot account for UnI-R's substantial improvements. Removing either ACDC or CFGSD reduces performance across all five benchmarks. In particular, removing ACDC means retain all hidden states from the future dyanmics and lowers BabyVision and VTBench scores by 3.09 and 2.58 points, respectively, highlighting the importance of selecting complementary future dynamics. Removing CFGSD leads to drops of 2.11 points on V* and 1.93 points on MME-RealWorld, supporting the value of contrastive guidance for reasoning with imagined futures. Together, these results identify dynamic state selection and contrastive future guidance as complementary contributors to UnI-R's gains beyond standard supervised fine tuning.

Ablation of the Loss Function. As shown in Table 4, removing any of the three auxiliary losses reduces performance across all five benchmarks. During CDFL, removing FutureNCE causes the largest degradation, highlighting the importance of discriminative future representations. Removing Self-Align also weakens performance, supporting the value of consistency across rollouts of the same input. During CFGSD, positive teacher matching alone underperforms the full contrastive objective, demonstrating the benefit of separating predictions induced by counterfactual futures. These results support the complementary roles of representation discrimination, rollout consistency, and contrastive future guidance in reasoning with imagined dynamics in unified latent space.

Hyperparameter Analysis. We independently vary the number of learnable queries R, readout blocks L_R, generated future frames T, and retained hidden states K, keeping other settings at their defaults. As shown in Figure 4, performance on benchmarks generally improves as readout capacity and the future information budget increase, then reaches a broad plateau. With sufficient capacity and future information, UnI-R maintains strong performance across a range of configurations, indicating its gains do not depend on a narrowly tuned setting. Adding more queries or blocks, generating more frames, or retaining more states provides limited additional benefit. We therefore adopt $R = 32$, $L_R = 3$, $T = 32$, and $K = 256$, capturing most observed gains without unnecessary increases in model capacity or future processing. These defaults provide a practical balance between reasoning performance and training and inference costs. Detailed results are in Appendix D.2.

Generalization across Model Scales and Architectures. We evaluate UnI-R on Cosmos3-Nano-8B, Cosmos3-Edge-2B, and Lance-3B in Table 5, observing consistent improvements across all five benchmarks on every backbone. On the smaller Cosmos3-Edge-2B, UnI-R doubles the BabyVision score from 5.41 to 10.82 and improves VTBench by 7.74 points, demonstrating its effectiveness beyond the default model scale. The benefits also extend to Lance-3B (Fu et al., 2026), with gains of 6.81 points on V* and 6.98 points on MME-RealWorld. These results show that UnI-R's improvements are not restricted to a particular parameter scale or the Cosmos3 architecture, supporting its broader applicability to unified models of understanding and generation.

Table 5: UnI-R across different backbones.

Model	BabyVision	VTBench	V*	MME-RW	SpatialEval
cosmos3-Nano-8B	8.51	23.87	75.92	44.40	64.69
+UnI-R	**17.27**	**31.29**	**81.16**	**52.74**	**68.54**
cosmos3-Edge-2B	5.41	19.03	70.68	39.66	50.31
+UnI-R	**10.82**	**26.77**	**78.01**	**47.11**	**54.84**
Lance-3B	5.67	14.84	60.73	35.28	42.78
+UnI-R	**10.05**	**20.32**	**67.54**	**42.26**	**45.90**

Table 6: UnI-R versus full attention.

Model	BabyVision	VTBench	V*	MME-RW	SpatialEval
Qwen3-VL-8B	7.47	23.22	73.82	44.71	64.05
Cosmos3-8B	8.51	23.87	75.92	44.40	64.69
+full attention	3.09	13.87	37.17	34.13	42.01
+full attention*	6.96	21.61	76.44	43.56	61.88
+UnI-R	**17.27**	**31.29**	**81.16**	**52.74**	**68.54**

* denotes training with the UnI-R data.

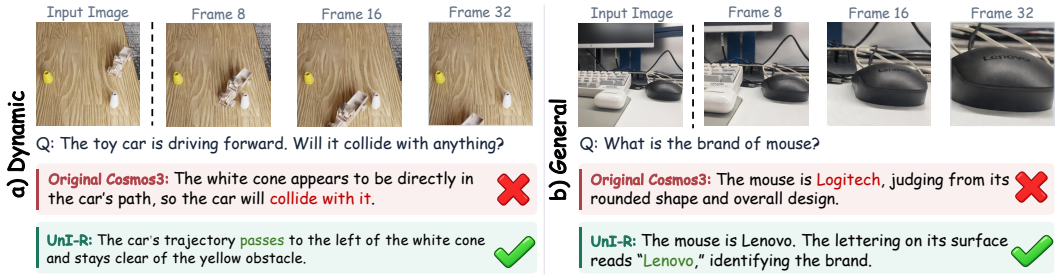


Figure 5: **Case studies of UnI-R.** Imagined futures support (a) dynamic reasoning by clarifying motion trajectories and (b) general visual understanding through closer inspection of relevant details.

Why Not Full Attention? Allowing the reasoning stream to attend directly to all generated future states offers a straightforward alternative to UnI-R. However, enabling full attention at inference time degrades performance across all benchmarks. As shown in Table 6, standard SFT on the same data as UnI-R recovers much of the lost performance but still underperforms the original Cosmos3. These results suggest that access to generated states alone is insufficient for effective foresight reasoning. The model must extract useful dynamics from stochastic and redundant future representations and learn to use them for reasoning. ACDC and F2R select and integrate complementary future dynamics into reasoning. CDFL promotes consistent, discriminative representations, while CFGSD grounds predictions through contrastive supervision from factual and counterfactual futures. These components support effective use of dynamics, beyond merely enabling access to generated states.

4.3 Case Study

Figure 5 illustrates how imagined futures can support dynamic reasoning and general visual understanding. For dynamic questions, simulated evolution links observed spatial configurations to motion trajectories and interaction outcomes, enabling judgments beyond static proximity. For general visual questions, imagined changes in viewing scale facilitate closer inspection of relevant regions, emphasizing informative local details over overall appearance. These cases highlight that foresight anticipates scene evolution and facilitates detailed visual inspection. By integrating latent dynamics into the reasoning stream, UnI-R extends the utility of world generation beyond future prediction to broader visual reasoning in unified latent space. More qualitative visualizations are in Appendix E.

5 Conclusion

We presented Understand-Imagine-Reason (UnI-R), a framework that connects future dynamics with visual reasoning in a unified latent space through a loop of understanding, imagination, and reasoning. ACDC selects complementary future dynamics, while F2R aggregates and integrates them into the reasoning stream. CDFL promotes consistency across rollouts and discrimination across inputs, while CFGSD grounds predictions through contrastive supervision from factual and counterfactual futures. These components enable the model to reason with internally imagined dynamics without access to real future observations at inference. Extensive experiments demonstrate substantial gains in both dynamic and general visual reasoning, with consistent improvements across model scales and unified architectures. These findings suggest that visual reasoning can benefit not only from interpreting the observed scene, but also from internally simulating how it may evolve.

AI USE STATEMENT

In preparing this manuscript, we used AI tools to assist with writing and language polishing, including improving grammar, wording, clarity, and organization. The authors reviewed all AI-assisted text, checking technical descriptions, mathematical notation, and statements about experimental results for accuracy and consistency with the underlying work. We take full responsibility for the final content of this paper, including all text, claims, and artifacts produced with the assistance of AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the UnI-R architecture and training objectives in Sections 3, with pseudocode for ACDC provided in Algorithm 1. Section 4 and Appendix C documents the evaluation benchmarks, comparison methods, and training and evaluation settings, including the trainable components, default hyperparameters, and decoding configuration.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023b.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Halely Balaban and Tomer D Ullman. The capacity limits of moving objects in the imagination. *Nature Communications*, 16(1):5899, 2025.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332, 2013.

Johannes Bill, Samuel J Gershman, and Jan Drugowitsch. Visual motion perception as online hierarchical inference. *Nature communications*, 13(1):7403, 2022.

Kartik Chandra, Tony Chen, Tzu-Mao Li, Jonathan Ragan-Kelley, and Josh Tenenbaum. Inferring the future by imagining the past. *Advances in Neural Information Processing Systems*, 36:21196–21216, 2023.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Yiping Bao, et al. Babyvision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

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? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, pp. 49250–49267, 2023.

Ingmar EJ de Vries and Moritz F Wurm. Predictive neural representations of naturalistic dynamic input. *Nature Communications*, 14(1):3858, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jason Fischer, John G Mikhael, Joshua B Tenenbaum, and Nancy Kanwisher. Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, 113 (34):E5072–E5081, 2016.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, et al. Lance: Unified multimodal modeling by multi-task synergy. *arXiv preprint arXiv:2605.18678*, 2026.

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.

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.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R Fung. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16427–16444, 2025.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhe Huang, Hao Wen, Aiming Hao, Bingze Song, Meiqi Wu, Jiahong Wu, Xiangxiang Chu, Sheng Lu, and Haoqian Wang. Taming hallucinations: Boosting mllms' video understanding via counterfactual video generation. *arXiv preprint arXiv:2512.24271*, 2025.

Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. Visual structures help visual reasoning: Addressing the binding problem in LVLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zoe Kourtzi and Nancy Kanwisher. Activation in human mt/mst by static images with implied motion. *Journal of Cognitive Neuroscience*, 12(1):48–55, 2000. doi: 10.1162/08989290051137594.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 71683–71702. Curran Associates, Inc., 2023.

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

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9098–9108, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. Deepscan: A training-free framework for visually grounded reasoning in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19155–19164, June 2026b.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, pp. 26763–26773, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024b.

Junhua Liu, Zhangcheng Wang, Zhike Han, Ningli Wang, Guotao Liang, and Kun Kuang. Twiff (think with future frames): A large-scale dataset for dynamic visual reasoning. *arXiv preprint arXiv:2602.10675*, 2026.

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, 2024c.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, S Kevin Zhou, Yijun Yang, Alan Yuille, and Jieneng Chen. Causalspatial: A benchmark for object-centric causal spatial reasoning. *arXiv preprint arXiv:2601.13304*, 2026.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pp. 508–524. Springer, 2020.

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.

Gabriel Herbert Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, 2025.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. In *CVPR*, 2026.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, et al. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*, 2026a.

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. In *CVPR*, 2026b.

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, 2025.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, GUANNAN JIANG, Xiaoshuai Sun, and Rongrong Ji. ControlMLLM: Training-free visual prompt learning for multimodal large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

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.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

xAI. Grok-1.5 vision preview. Technical report, 2024.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *CVPR*, 2026.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning*, 2024.

Xinyu Zhang, Yuxuan Dong, Lingling Zhang, Chengyou Jia, Zhuohang Dang, Basura Fernando, Jun Liu, and Mike Zheng Shou. CoFFT: Chain of foresight-focus thought for visual language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

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.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. FOCUS: Internal MLLM representations for efficient fine-grained visual question answering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Yucheng Zhou, Wei Tao, Yiwen Guo, and Jianbing Shen. World models meet language models: On the complementarity of concrete and abstract reasoning. *arXiv preprint arXiv:2606.03603*, 2026.

A Appendix Overview

This appendix provides additional theoretical analysis, implementation details, and experimental results for UnI-R. Appendix B presents the theoretical analysis of ACDC, establishing the monotonicity and submodularity of its coverage objective and the approximation guarantee of its greedy selection procedure. Appendix C describes the evaluation benchmarks, comparison methods, and detailed training and evaluation settings. Appendix D reports extended benchmark results and detailed hyperparameter ablations, covering the number of learnable queries, readout Transformer blocks, generated future frames, and retained tokens. Appendix E presents additional qualitative case studies illustrating how imagined future dynamics support visual reasoning. Appendix F discusses the limitations of UnI-R and directions for future work.

B Derivations and Proofs

B.1 Theoretical Analysis of ACDC

We analyze the anchor-conditioned coverage objective in Eq. equation 3 and the greedy procedure in Algorithm 1. Throughout this subsection, the anchor set $\mathcal{A}$ is finite and nonempty, the candidate set $\mathcal{V}$ is finite, and all features, positions, and similarities are held fixed during one execution of the selection algorithm. The budget K is an integer satisfying $0 \le K \le |\mathcal{V}|$. For brevity, write $f(\mathcal{S}) = F(\mathcal{S} \mid \mathcal{A})$. The analysis concerns selection for a fixed model and a fixed collection of generated futures.

Kernel interpretation and residual coverage. Assume that every projected feature used in the kernel is nonzero, so that $\|z_i\|_2 = 1$ for all $i \in \mathcal{A} \cup \mathcal{V}$. The similarity in Equation 2 can be written as

$$\begin{aligned} k(i,j) &= \exp\big(z_i^\top z_j - 1 - \|p_i - p_j\|_2^2\big) \\ &= \exp\left(-\frac{1}{2}\|z_i - z_j\|_2^2 - \|p_i - p_j\|_2^2\right), \end{aligned} \tag{14}$$

where the second equality follows from $\|z_i - z_j\|_2^2 = 2 - 2z_i^\top z_j$. Thus, $0 < k(i,j) \le 1$ and $k(i,i) = 1$. The kernel measures proximity in the projected feature and spatial coordinate spaces. For each candidate $i \in \mathcal{V}$, define its current coverage under a selected set $\mathcal{S} \subseteq \mathcal{V}$ by

$$c_i(\mathcal{S}) = \max_{j \in \mathcal{A} \cup \mathcal{S}} k(i,j). \tag{15}$$

Since the anchors are always included when computing coverage, $c_i(\mathcal{S}) \ge b_i$, where $b_i = \max_{a \in \mathcal{A}} k(i,a)$. These maxima are well-defined because their indexing sets are finite and nonempty. Define the nonnegative residual similarity

$$w_{ij} = [k(i,j) - b_i]_+, \quad [x]_+ = \max\{0, x\}. \tag{16}$$

The objective admits the equivalent form

$$f(\mathcal{S}) = \sum_{i \in \mathcal{V}} \big(c_i(\mathcal{S}) - b_i\big) = \sum_{i \in \mathcal{V}} \max_{j \in \mathcal{S}} w_{ij}, \tag{17}$$

where the maximum over an empty set of residual similarities is defined to be zero. Indeed, for a nonempty $\mathcal{S}$, subtracting b_i from $\max\{b_i, \max_{j\in\mathcal{S}} k(i,j)\}$ gives $\max_{j\in\mathcal{S}}[k(i,j)-b_i]_+$; the empty-set case follows from $c_i(\varnothing) = b_i$. Consequently, the ACDC objective is a facility-location objective over residual similarities beyond the anchor coverage. The kernel bounds also give

$$0 \le f(\mathcal{S}) \le \sum_{i\in\mathcal{V}} (1 - b_i) = f(\mathcal{V}). \tag{18}$$

The last equality holds because selecting all candidates includes candidate i itself, whose similarity to itself is one.

Exact marginal gain and coverage updates. For $j \in \mathcal{V} \setminus \mathcal{S}$, the marginal gain of adding j is

$$\begin{aligned} \Delta(j \mid \mathcal{S}) &= f(\mathcal{S}\cup\{j\}) - f(\mathcal{S}) \\ &= \sum_{i\in\mathcal{V}} [\max\{c_i(\mathcal{S}), k(i,j)\} - c_i(\mathcal{S})] \\ &= \sum_{i\in\mathcal{V}} [k(i,j) - c_i(\mathcal{S})]_+. \end{aligned} \tag{19}$$

This is exactly the score used in Algorithm 1. The maintained coverage vector is also exact. Initially, $\mathcal{S}_0 = \varnothing$ and $c_i^{(0)} = b_i = c_i(\varnothing)$. Suppose $c_i^{(r)} = c_i(\mathcal{S}_r)$. After selecting $j_r \in \mathcal{V} \setminus \mathcal{S}_r$ and setting $\mathcal{S}_{r+1} = \mathcal{S}_r \cup \{j_r\}$, the update gives

$$c_i^{(r+1)} = \max\{c_i^{(r)}, k(i,j_r)\} = \max_{j\in\mathcal{A}\cup\mathcal{S}_{r+1}} k(i,j) = c_i(\mathcal{S}_{r+1}). \tag{20}$$

Induction therefore shows that the maintained coverage vector and the resulting marginal gains are exact at every iteration.

Proposition 1: Normalization, monotonicity, and submodularity. For fixed anchors and similarities, f is a normalized, nonnegative, monotone submodular set function on $\mathcal{V}$.

Proof. The identity $c_i(\varnothing) = b_i$ gives $f(\varnothing) = 0$. Equation 19 shows that every marginal gain is nonnegative, establishing monotonicity and, together with normalization, nonnegativity.

To prove submodularity, take $\mathcal{S} \subseteq \mathcal{U} \subseteq \mathcal{V}$ and $j \in \mathcal{V} \setminus \mathcal{U}$. The inclusion of the selected sets implies $c_i(\mathcal{S}) \le c_i(\mathcal{U})$ for every candidate i. Hence,

$$\Delta(j \mid \mathcal{S}) = \sum_{i\in\mathcal{V}} [k(i,j) - c_i(\mathcal{S})]_+ \ge \sum_{i\in\mathcal{V}} [k(i,j) - c_i(\mathcal{U})]_+ = \Delta(j \mid \mathcal{U}). \tag{21}$$

This is the diminishing-returns condition for submodularity: a candidate cannot provide more additional coverage after the selected set has grown. □

Proposition 2: Greedy approximation guarantee. Let $1 \le K \le |\mathcal{V}|$, and let

$$\mathcal{S}^\star \in \operatorname*{argmax}_{\mathcal{S}\subseteq\mathcal{V},\ |\mathcal{S}|\le K} f(\mathcal{S}). \tag{22}$$

Suppose Algorithm 1 starts from $\mathcal{S}_0 = \varnothing$ and, at each of K iterations, adds an unselected candidate attaining the largest exact marginal gain. Its output $\mathcal{S}_K$ then satisfies

$$f(\mathcal{S}_K) \ge \left[1 - \left(1 - \frac{1}{K}\right)^K\right] f(\mathcal{S}^\star) \ge (1 - e^{-1}) f(\mathcal{S}^\star). \tag{23}$$

Ties between candidates with equal maximal marginal gains may be resolved arbitrarily. This is the classical greedy guarantee for monotone submodular maximization under a cardinality constraint (Nemhauser et al., 1978). We give the argument specialized to ACDC below.

Proof. Consider iteration $r < K$. By monotonicity, adding $\mathcal{S}_r$ to the optimal set cannot decrease its value. Next, add the elements of $\mathcal{S}^\star \setminus \mathcal{S}_r$ to $\mathcal{S}_r$ one at a time. Submodularity bounds each resulting increment by that element's marginal gain at $\mathcal{S}_r$. Therefore,

$$f(\mathcal{S}^\star) - f(\mathcal{S}_r) \le f(\mathcal{S}^\star \cup \mathcal{S}_r) - f(\mathcal{S}_r) \le \sum_{j\in\mathcal{S}^\star\setminus\mathcal{S}_r} \Delta(j \mid \mathcal{S}_r) \le K\, \Delta(j_r \mid \mathcal{S}_r). \tag{24}$$

The final inequality uses $|\mathcal{S}^\star \setminus \mathcal{S}_r| \le K$, the nonnegativity of marginal gains, and the maximality of the greedy marginal gain. Since $\Delta(j_r \mid \mathcal{S}_r) = f(\mathcal{S}_{r+1}) - f(\mathcal{S}_r)$, we obtain

$$f(\mathcal{S}^\star) - f(\mathcal{S}_{r+1}) \le \left(1 - \frac{1}{K}\right) [f(\mathcal{S}^\star) - f(\mathcal{S}_r)]\,. \tag{25}$$

Starting from $f(\mathcal{S}_0) = 0$ and applying this inequality for K iterations yields

$$f(\mathcal{S}^\star) - f(\mathcal{S}_K) \le \left(1 - \frac{1}{K}\right)^K f(\mathcal{S}^\star). \tag{26}$$

Rearranging gives the first inequality in Equation 23. The second follows from $(1 - 1/K)^K \le e^{-1}$. For $K = 0$, the only feasible set is the empty set, and both the greedy and optimal objective values are zero. □

Consequences for anchor conditioning and redundancy. Equation 19 implies the exact zero-gain criterion

$$\Delta(j \mid \mathcal{S}) = 0 \quad \Longleftrightarrow \quad k(i,j) \le c_i(\mathcal{S}) \quad \text{for every } i \in \mathcal{V}. \tag{27}$$

Thus, a candidate receives no additional reward when its entire coverage profile is already explained by the anchors and selected candidates.

For the particular kernel in Equation 14, this criterion admits a more explicit characterization. Because the marginal-gain sum includes $i = j$,

$$\Delta(j \mid \mathcal{S}) \ge [k(j,j) - c_j(\mathcal{S})]_+ = 1 - c_j(\mathcal{S}). \tag{28}$$

Consequently, zero marginal gain requires $c_j(\mathcal{S}) = 1$. Since $\mathcal{A} \cup \mathcal{S}$ is finite and nonempty, there must then exist $\ell \in \mathcal{A} \cup \mathcal{S}$ such that $k(j,\ell) = 1$. By Equation 14, this is equivalent to $z_j = z_\ell$ and $p_j = p_\ell$. Conversely, these equalities imply $k(i,j) = k(i,\ell) \le c_i(\mathcal{S})$ for every $i \in \mathcal{V}$, which gives zero marginal gain. Therefore,

$$\Delta(j \mid \mathcal{S}) = 0 \quad \Longleftrightarrow \quad \exists\, \ell \in \mathcal{A} \cup \mathcal{S} : (z_j, p_j) = (z_\ell, p_\ell). \tag{29}$$

In particular, a candidate that duplicates the projected feature and spatial position of an anchor receives zero marginal gain whenever it remains unselected. The same conclusion holds when it duplicates the comparison representation of an already selected candidate. This characterizes exact redundancy in the feature-position space used by ACDC. Near-duplicate candidates may still have positive marginal gains, and a fixed-budget implementation may select zero-gain candidates after all positive marginal gains have been exhausted.

C Detailed Experimental Settings

Appendix C.1 introduces the benchmarks used in our evaluation, Appendix C.2 describes the compared methods, and Appendix C.3 provides detailed hyperparameter settings.

C.1 Benchmarks

We evaluate UnI-R on two categories of benchmarks. *Dynamic-Centric Image Understanding Tasks* include BabyVision (Chen et al., 2026), Visual Cognitive Tasks (Schulze Buschoff et al., 2025), VTBench (Tong et al., 2026), OpenWorldQA (Zhou et al., 2026), and CausalSpatial (Ma et al., 2026). These benchmarks emphasize physical prediction, causal inference, spatial transformations, and visually grounded reasoning. *General Image Understanding Tasks* include MMBench (Liu et al., 2024c), MMStar (Chen et al., 2024a), MM-Vet (Yu et al., 2024), V* (Wu & Xie, 2024), HRBench4K/8K (Wang et al., 2025), MME-RealWorld (Zhang et al., 2025), and SpatialEval (Wang et al., 2024). They assess broader capabilities such as object recognition, text reading, fine-grained perception, and spatial understanding. Together, these benchmarks allow us to examine whether UnI-R supports reasoning about scene dynamics while also benefiting general visual understanding.

BabyVision (Chen et al., 2026). BabyVision evaluates visual reasoning that relies on precise perception rather than extensive linguistic knowledge. It contains questions across 22 subtypes, organized into four categories: Fine-Grained Discrimination, Visual Tracking, Spatial Perception, and Visual Pattern Recognition. Representative tasks include identifying subtle visual differences, tracing

paths, solving mazes, interpreting three-dimensional views, unfolding cubes, folding paper, counting blocks, and completing visual patterns. These tasks require models to preserve local visual details while reasoning about spatial structure and transformations. BabyVision therefore complements physical prediction tasks by assessing perceptual and spatial reasoning abilities that cannot be adequately evaluated through semantic recognition alone.

Visual Cognitive Tasks (Schulze Buschoff et al., 2025). Visual Cognitive Tasks probe intuitive physics, causal reasoning, and intuitive psychology through controlled visual scenarios. We focus on four representative tasks: Block Tower, Michotte, Help or Hinder, and Astronaut. Block Tower asks models to judge the stability of photographed wooden block structures, requiring reasoning about support and balance. Michotte examines two-ball interactions, including collision outcomes and counterfactual trajectories when one ball is absent. Astronaut requires models to infer movement costs and goal rewards from an agent's routes and choices. Help or Hinder evaluates whether models can infer how one agent facilitates or obstructs another agent's goal-directed behavior. Together, these tasks assess physical prediction, counterfactual reasoning, and the interpretation of goal-directed behavior.

VTBench (Tong et al., 2026). We use VTBench to denote the Eyeballing Puzzles and Visual Puzzles subsets of VideoThinkBench. Eyeballing Puzzles assesses geometric intuition through tasks involving point placement, line construction, and relationships among geometric shapes. Visual Puzzles evaluates inductive reasoning over structured patterns, requiring models to identify rules involving attributes such as color, shape, and size, and then infer a missing element or complete a pattern. These subsets provide complementary assessments of geometric perception and abstract visual reasoning.

OpenWorldQA (Zhou et al., 2026). OpenWorldQA evaluates physical future prediction in real-world scenes from a single observed image. Its questions are constructed from short videos, but the evaluated model receives only an anchor frame preceding the outcome, together with the question and answer options. The benchmark contains human-verified, four-choice questions spanning 12 physical reasoning categories. Question types cover event order, counting, first contact, intermediate states, failure, and counterfactual outcomes. Since ground-truth future frames are withheld, models must infer subsequent events from the initial visual conditions rather than recognize an already observed outcome. This setting directly assesses whether UnI-R can anticipate scene evolution and the consequences of physical interactions.

CausalSpatial (Ma et al., 2026). CausalSpatial evaluates object-centric causal spatial reasoning in three-dimensional scenes. Given an initial scene image and a question describing hypothetical object motion or placement, models must predict the resulting spatial consequences. The benchmark covers four categories: Collision, Compatibility, Occlusion, and Trajectory. Collision concerns whether moving objects will interact, while Compatibility examines whether objects can fit within or be placed in specified spatial configurations. Occlusion tests how changes in object position affect visibility, and Trajectory assesses motion resulting from physical interactions. Two difficulty levels distinguish interactions between a pair of objects from those involving a target and multiple surrounding objects. CausalSpatial therefore tests whether models can connect scene geometry with the consequences of hypothetical changes, rather than merely identify existing spatial relationships.

MMBench (Liu et al., 2024c). MMBench evaluates general multimodal understanding through multiple-choice questions covering 20 fine-grained ability dimensions, including object localization, attribute recognition, action understanding, OCR, and relation reasoning. It provides both English and Chinese versions, supporting evaluation across languages. Its CircularEval protocol examines prediction consistency under cyclic permutations of answer options, reducing the influence of option-order bias on evaluation. The diversity of its visual content and reasoning requirements makes MMBench useful for assessing whether improvements extend across multiple aspects of image understanding rather than remaining confined to a specific task family.

MMStar (Chen et al., 2024a). MMStar is designed to assess multimodal abilities while limiting reliance on language priors and textual shortcuts. It contains selected questions organized into six core capabilities and 18 fine-grained dimensions. The six capabilities are coarse perception, fine-grained perception, instance reasoning, logical reasoning, science and technology, and mathematics, with 250 questions per capability. Its curation process prioritizes questions that require visual evidence

and filters examples that can be answered without inspecting the image. This design helps assess whether improvements arise from more effective use of visual information.

MM-Vet (Yu et al., 2024). MM-Vet evaluates the integration of multiple vision-language capabilities through open-ended questions. It covers six dimensions: recognition, OCR, knowledge, language generation, spatial awareness, and mathematics. Rather than evaluating these capabilities in isolation, its questions often require several of them to be combined, such as reading text while interpreting spatial relationships and applying relevant knowledge. The benchmark uses an LLM-based evaluation procedure to score free-form responses across different answer formats. MM-Vet thus complements multiple-choice benchmarks by assessing whether models can produce accurate answers to questions requiring coordinated perception and reasoning.

V* (Wu & Xie, 2024). V*Bench, denoted as V* in our experiments, evaluates fine-grained visual understanding in high-resolution images. It focuses on attribute recognition and spatial relationship reasoning, with questions that depend on small or easily overlooked visual details. Answering these questions requires identifying the relevant objects and inspecting localized evidence that may not be captured by a global image representation. This benchmark assesses whether UnI-R can retain and use task-relevant visual details when the answer depends on precise grounding rather than recognition of the overall scene.

HRBench (Wang et al., 2025). HRBench evaluates high-resolution image understanding under two settings, HRBench4K and HRBench8K. Its tasks cover fine-grained perception of individual objects and relationships across multiple objects, emphasizing both local attributes and spatial configurations. The 4K setting is constructed by cropping regions containing question-relevant objects from the corresponding 8K images. These settings expose the difficulty of recognizing small details and relating distant visual evidence in large images, where downsampling can remove information needed to answer the question. We include both settings to assess UnI-R under different image resolutions and spatial extents.

MME-RealWorld (Zhang et al., 2025). MME-RealWorld evaluates multimodal understanding in complex, high-resolution real-world environments. It contains manually annotated questions covering 43 subtasks across five domains: OCR in the wild, remote sensing, diagrams and tables, autonomous driving, and monitoring. Its images have an average resolution of approximately 2000×1500, and many questions require identifying small objects, reading detailed text, or interpreting relationships among multiple visual elements. The benchmark combines fine-grained perception with application-specific reasoning, allowing us to assess whether UnI-R remains effective across diverse visual environments beyond controlled physical and geometric scenarios.

RealWorldQA. RealWorldQA is a visual question-answering benchmark introduced by xAI to evaluate the spatial understanding of multimodal models in real-world scenes. It contains images paired with a question and a verifiable answer, including anonymized images captured from vehicles and other everyday settings. Questions cover relative positions, object orientation, size and distance comparisons, counting, and navigation-related judgments. The benchmark includes both multiple-choice and short-answer questions, requiring models to interpret visual evidence within its surrounding spatial context. By emphasizing relationships among objects and their environment rather than object recognition alone, RealWorldQA complements general image understanding benchmarks with a focused assessment of visually grounded spatial reasoning.

SpatialEval (Wang et al., 2024). SpatialEval evaluates spatial reasoning across four complementary dimensions: spatial relationships, positional understanding, object counting, and navigation. It comprises four task settings, Spatial-Map, Maze-Nav, Spatial-Grid, and Spatial-Real, combining structured synthetic environments with natural images. These tasks require models to interpret object layouts, distinguish relative positions, count spatially arranged objects, and reason about routes through an environment. SpatialEval provides a focused assessment of whether models can recover and reason over spatial structure across different visual representations, complementing the broader capability coverage of general multimodal benchmarks.

C.2 Other Methods

We compare UnI-R with methods from three paradigms, namely *Thinking About Images*, *Thinking With Images*, and *Latent Visual Reasoning*. These methods improve visual reasoning through textual

reasoning, explicit visual evidence manipulation, or continuous intermediate representations. Below, we summarize their computational procedures and training objectives using unified notation.

C.2.1 THINKING ABOUT IMAGES

Vision-R1 (Huang et al., 2026) develops multimodal textual reasoning through cold-start initialization followed by reinforcement learning. Its modality bridging procedure converts image information into detailed textual descriptions, which support the generation of reasoning traces by a text-based reasoning model. Filtered traces are paired with the original images for supervised training:

$$\mathcal{L}_{\text{cold}} = -\mathbb{E}_{(I,q,o)\sim\mathcal{D}_{\text{cold}}}\left[\sum_{t=1}^{|o|}\log \pi_\theta(o_t \mid I, q, o_{<t})\right], \tag{30}$$

where o contains the textual reasoning and answer.

The subsequent Progressive Thinking Suppression Training initially constrains output length and progressively relaxes this restriction. At stage s, outputs are sampled from

$$\mathcal{O}^{(s)} = \{o : |o| \le L_s\}, \qquad L_{s+1} \ge L_s. \tag{31}$$

Its hard formatting result reward requires both answer correctness and valid formatting:

$$R(o) = \mathbf{1}\left[\text{Correct}(o) \wedge \text{ValidFormat}(o)\right]. \tag{32}$$

For a group of G sampled outputs, GRPO uses normalized rewards and policy ratios:

$$\hat{A}_i = \frac{R(o_i) - \text{mean}_j\, R(o_j)}{\text{std}_j\, R(o_j)}, \qquad \rho_i(\theta) = \frac{\pi_\theta(o_i \mid I, q)}{\pi_{\theta_{\text{old}}}(o_i \mid I, q)}. \tag{33}$$

The clipped objective is

$$\mathcal{J}_{\text{GRPO}} = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i(\theta)\hat{A}_i, \\ \text{clip}\left(\rho_i(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\Big) - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)\Bigg]. \tag{34}$$

Here, ϵ controls clipping and β controls reference-policy regularization. This training procedure refines verbalized reasoning while gradually permitting longer reasoning trajectories.

ViGoRL (Sarch et al., 2026) grounds textual reasoning in explicit image coordinates. Each intermediate step combines a natural-language thought with a spatial reference:

$$n_t = \langle s_t, (x_t, y_t)\rangle, \qquad \tau = (n_1, \ldots, n_T, \hat{y}), \tag{35}$$

where s_t describes the reasoning operation and (x_t, y_t) identifies its associated image location. The trajectory probability factorizes as

$$\pi_\theta(\tau \mid I, q) = \left[\prod_{t=1}^{T}\pi_\theta(n_t \mid I, q, n_{<t})\right]\pi_\theta(\hat{y} \mid I, q, n_{\le T}). \tag{36}$$

A supervised warm-start stage uses teacher-generated trajectories constructed through Monte Carlo Tree Search, including successful reasoning chains and corrected trajectories that backtrack from unsuccessful steps.

ViGoRL subsequently applies GRPO with task and formatting rewards:

$$R(\tau) = \lambda_{\text{fmt}} r_{\text{fmt}}(\tau) + \lambda_{\text{task}} r_{\text{task}}(\tau). \tag{37}$$

The formatting reward checks the required reasoning and answer structure together with the validity of coordinate references, while the task reward evaluates the answer. Group-normalized rewards guide policy updates toward successful grounded trajectories. The coordinates serve as explicit references within the generated text, linking intermediate reasoning to particular image locations rather than leaving visual references entirely implicit.

C.2.2 THINKING WITH IMAGES

ControlMLLM (Wu et al., 2024a) introduces visual prompts into a frozen LVLM through test-time optimization of its projected visual tokens. Given visual embeddings

$$V = \mathrm{Proj}\left(\mathrm{Enc}_v(I)\right) \in \mathbb{R}^{N\times d}, \tag{38}$$

it introduces an additive latent variable Δ:

$$V^{\mathrm{ctrl}} = V + \Delta, \qquad \Delta^{(0)} = 0. \tag{39}$$

For a supplied region mask $\mathcal{R}$, let $\bar{A}_i(\Delta)$ denote the aggregated attention from textual prompt tokens to visual token i. The hard-mask attention objective is

$$\mathcal{E}(\Delta;\mathcal{R}) = \left(1 - \frac{\sum_{i\in\mathcal{R}} \bar{A}_i(\Delta)}{\sum_{i=1}^{N} \bar{A}_i(\Delta)}\right)^2. \tag{40}$$

The latent variable is updated while model parameters remain fixed:

$$\Delta^{(k+1)} = \Delta^{(k)} - \eta\nabla_{\Delta}\mathcal{E}(\Delta^{(k)};\mathcal{R}), \tag{41}$$

where η is the optimization step size. Answer generation then conditions on the modified visual embeddings:

$$\hat{y} = f_\theta(V + \Delta^*, q). \tag{42}$$

The method also supports soft visual prompts through distance-based masks. Its central operation is therefore attention-guided adjustment of the visual input representation, rather than model fine-tuning or generation of an intermediate reasoning image.

DyFo (Li et al., 2025) performs training-free visual search through a Focus Adjuster and Focus Tree Search. A focus state contains a visual region and its semantic description:

$$f_t = (I_t, T_t). \tag{43}$$

Given an action a_t, an LVLM $\mathcal{L}$ updates the semantic focus, and a visual expert $\mathcal{E}$ obtains the corresponding image region:

$$\begin{aligned} T_{t+1} &= \mathcal{L}(f_t, a_t), \\ I_{t+1} &= \mathcal{E}(T_{t+1}, I_t, a_t), \\ f_{t+1} &= (I_{t+1}, T_{t+1}). \end{aligned} \tag{44}$$

Semantic Focus narrows attention toward relevant objects, whereas Semantic Scatter expands the visual context.

Focus Tree Search explores these alternatives using Monte Carlo Tree Search. Its branch selection balances estimated value and exploration:

$$a^* = \arg\max_{a\in\mathcal{A}(f)} \left[Q(f,a) + w\sqrt{\frac{\log N(f)}{N(c(f,a))}}\right], \tag{45}$$

where $c(f,a)$ is the child state, N denotes visit counts, and w controls exploration. Focus rewards account for visual-semantic consistency and effective focus area. Answers from explored nodes are combined through weighted voting:

$$\hat{y} = \arg\max_{y} \sum_{f\in\mathcal{F}} R_f\, \mathbf{1}[\hat{y}_f = y]. \tag{46}$$

Here, $\mathcal{F}$ contains the explored focus states, $\hat{y}_f$ is a node-level answer, and R_f is its reward. The final prediction thus integrates evidence obtained at different visual granularities.

CoFFT (Zhang et al., 2026) alternates between candidate reasoning generation, Dual Foresight Decoding, and Visual Focus Adjustment. Given reasoning history R_t and focused image I_t, it samples candidate continuations containing multiple reasoning steps.

Let A_q and A_s denote relative attention maps associated with the question and candidate s. Candidates receive visual-focus and reasoning-progression scores:

$$\begin{aligned} E_{\text{att}}(s) &= \frac{1}{2}\cos(A_q, A_s) + \frac{1}{2}\,\text{IoU}_{30\%}(A_q, A_s), \\ E_{\text{prob}}(s) &= \frac{1}{\ell}\sum_{j=1}^{\ell}(p_j - p_0), \end{aligned} \tag{47}$$

where $\text{IoU}_{30\%}$ measures overlap between high-attention regions, and p_j is the mean log probability after appending j candidate steps. Scores are normalized across candidates and combined:

$$E(s) = \lambda\,\text{Softmax}_{\mathcal{S}}\left(E_{\text{att}}\right)_s + (1-\lambda)\,\text{Softmax}_{\mathcal{S}}\left(E_{\text{prob}}\right)_s. \tag{48}$$

Only the first step of the highest-scoring continuation is committed:

$$s^* = \arg\max_{s\in\mathcal{S}} E(s), \qquad R_{t+1} = [R_t; \text{FirstStep}(s^*)]. \tag{49}$$

The next visual focus is guided by

$$\begin{aligned} C_t &= \max(A_q - \alpha A_{R_t}, 0), \\ A_{\text{crop}} &= \frac{1}{2}C_t + \frac{1}{2}A_{s^*}. \end{aligned} \tag{50}$$

A sliding window identifies the region with the highest mean score:

$$B^* = \arg\max_{B\in\Omega} \frac{1}{|B|}\sum_{u\in B} A_{\text{crop}}(u). \tag{51}$$

The crop is adopted when its score exceeds the global mean by an adaptive margin. Otherwise, the global view is retained. Its foresight concerns subsequent reasoning steps rather than future scene states.

DeepScan (Li et al., 2026b) combines Hierarchical Scanning, Refocusing, and Evidence-Enhanced Reasoning. Hierarchical Scanning first searches image patches for local cues. Within a connected relevance component G_k, it selects an interior point using

$$c_k = \arg\max_{u\in G_k} \widetilde{A}(u)\widetilde{d}(u, \partial G_k), \tag{52}$$

where $\widetilde{A}$ is normalized relevance and $\widetilde{d}$ is normalized distance from the component boundary. These points are mapped to the original image and used for segmentation-based evidence extraction.

Refocusing adjusts the context surrounding the recovered evidence. Starting from a merged evidence view V_1, it constructs

$$V_2 = \text{In}(V_1, q), \qquad V_3 = \text{Out}(V_1, s), \qquad V_4 = \text{In}(V_3, q), \tag{53}$$

where In crops around detected relevant content and Out expands the view by scale s. Candidate views are scored by

$$R(V) = \mathbf{1}_{V\rightsquigarrow q}\frac{HW}{h_V w_V}, \qquad V^* = \arg\max_{V\in\mathcal{V}} R(V). \tag{54}$$

The indicator records whether the LVLM judges the view sufficient for answering, while the area ratio favors compact evidence-preserving views.

A Hybrid Evidence Memory combines local evidence images $\{e_j\}_{j=1}^{M}$ with the selected contextual view:

$$\mathcal{H} = [e_1, \ldots, e_M, V^*], \qquad \hat{y} = f_\theta(\mathcal{H}, q). \tag{55}$$

The final prediction therefore uses both detailed local observations and a consolidated view of their surrounding context.

FOCUS (Zhong et al., 2026) uses internal LVLM representations to localize question-relevant objects without additional training. After extracting a target phrase from the question, it computes

similarities between its textual tokens and image tokens. For textual value vector $u_j^{(l)}$ and visual value vector $v_i^{(l)}$ at layer l,

$$A_j^{(l)}(i) = \frac{\left\langle u_j^{(l)}, v_i^{(l)} \right\rangle}{\left\| u_j^{(l)} \right\|_2 \left\| v_i^{(l)} \right\|_2}. \tag{56}$$

Layer-wise maps are aggregated and normalized. The maps associated with the target phrase tokens are then combined through elementwise multiplication:

$$A = \text{Normalize}\left(A_1 \odot A_2 \odot \cdots \odot A_m\right), \tag{57}$$

where m is the number of target tokens.

High-relevance locations provide anchors for adaptively expanded candidate regions. After overlapping proposals are filtered, candidates are ranked using the model's object-existence confidence. For a single target instance,

$$B^* = \arg\max_{B \in \mathcal{B}} S_{\text{exist}}(B, o), \tag{58}$$

where o is the target object and $\mathcal{B}$ is the candidate set. Multiple instances can retain several qualifying regions. The selected regions provide focused visual inputs for answering. Its localization signal comes from similarities between internal value representations, rather than conventional query-key attention alone.

VISER (Izadi et al., 2025) addresses visual binding errors through lightweight image structuring. These errors arise when a model identifies visual attributes but associates them with the wrong objects. Given an image $I \in \mathbb{R}^{H \times W \times 3}$, VISER applies a scaffold operator that adds horizontal lines and optional row indices:

$$I^{\text{vis}} = \mathcal{S}_n(I). \tag{59}$$

The n lines introduce $n+1$ ordered image regions,

$$\mathcal{R} = \{R_1, \ldots, R_{n+1}\}, \tag{60}$$

which serve as spatial references rather than independently processed image crops.

The transformed image is paired with a sequential scanning instruction:

$$q^{\text{vis}} = [q; p_{\text{scan}}], \tag{61}$$

where p_{scan} asks the model to inspect the image according to the horizontal structure. The answer is generated through the unchanged LVLM:

$$\hat{y} = f_\theta(I^{\text{vis}}, q^{\text{vis}}), \tag{62}$$

with the usual autoregressive factorization

$$p_\theta(y \mid I^{\text{vis}}, q^{\text{vis}}) = \prod_{t=1}^{|y|} p_\theta\left(y_t \mid y_{<t}, I^{\text{vis}}, q^{\text{vis}}\right). \tag{63}$$

The scaffold provides explicit spatial anchors, while the instruction encourages ordered inspection of the image. Importantly, this ordering is induced through the visual input and prompt rather than enforced by a separate recurrent module or additional model parameters. VISER therefore changes how visual evidence is presented to the model while leaving its architecture and weights unchanged.

ICoT (Gao et al., 2025) extends textual chain-of-thought reasoning by inserting visual rationales into the decoding sequence. Instead of describing all intermediate evidence in language, it alternates between selected visual tokens and textual reasoning:

$$\left(V_1^{\text{sel}}, r_1\right), \ldots, \left(V_T^{\text{sel}}, r_T\right), \hat{y}, \tag{64}$$

where V_t^{sel} contains visual evidence used by the following textual rationale r_t.

Let the image encoder and projector produce model-compatible visual tokens:

$$V = \text{Proj}\left(\text{Enc}_v(I)\right) = \{v_i\}_{i=1}^N, \qquad v_i \in \mathbb{R}^d. \tag{65}$$

ICoT uses Attention-driven Selection to identify relevant tokens when the model generates a predefined signal token. Let A_t denote the aggregated attention from that signal token to the original visual tokens. The selected indices are

$$\mathcal{I}_t = \mathrm{TopK}(A_t, K). \tag{66}$$

The corresponding tokens are restored to their original spatial order:

$$V_t^{\mathrm{sel}} = \mathrm{SpatialOrder}\left(\{v_i : i \in \mathcal{I}_t\}\right). \tag{67}$$

This ordering preserves the spatial arrangement of the selected evidence rather than arranging tokens by attention rank.

If C_t denotes the current embedding-level decoding context, the selected visual tokens are appended before textual generation resumes:

$$C_t^+ = \mathrm{Concat}\left(C_t, V_t^{\mathrm{sel}}\right), \qquad r_t \sim p_\theta(\cdot \mid C_t^+). \tag{68}$$

The resulting text is subsequently incorporated into the context for the next reasoning step. Through repeated selection and insertion, ICoT allows later reasoning to attend directly to localized evidence from the input image. The procedure requires neither parameter updates nor generation of new images, since its visual rationales are retrieved from the original visual token sequence.

C.2.3 Latent Visual Reasoning

CoVT (Qin et al., 2025) introduces continuous visual thought tokens supervised by multiple pretrained vision experts. These tokens encode complementary perceptual information, including segmentation, depth, edges, and DINO features. For expert type

$$m \in \mathcal{M} = \{\mathrm{seg}, \mathrm{depth}, \mathrm{edge}, \mathrm{dino}\}, \tag{69}$$

the LVLM autoregressively generates a corresponding token group:

$$Z^m = \{z_1^m, \ldots, z_{K_m}^m\}, \qquad z_k^m \in \mathbb{R}^d. \tag{70}$$

The training supervision requires these compact tokens to recover the corresponding perceptual information. For dense prediction tasks, the expert-specific reconstruction procedures can be summarized as

$$\hat{O}^m = \mathcal{D}_m\left(Z^m, F_m(I)\right), \qquad m \in \{\mathrm{seg}, \mathrm{depth}, \mathrm{edge}\}, \tag{71}$$

where $F_m(I)$ denotes image features used by the corresponding reconstruction procedure, and $\mathcal{D}_m$ includes its token projection and task-specific decoding operations. These features are distinct from the supervision targets $O^{m,*}$.

Segmentation tokens condition mask prediction and are supervised through matched segmentation targets. Depth and edge tokens support reconstruction of the respective dense maps. DINO supervision instead aligns projected visual thoughts with target patch features:

$$\hat{F}^{\mathrm{dino}} = \Pi_{\mathrm{dino}}(Z^{\mathrm{dino}}), \qquad \mathcal{L}_{\mathrm{dino}} = \left\| \hat{F}^{\mathrm{dino}} - F^{\mathrm{dino},*} \right\|_2^2. \tag{72}$$

The overall training objective combines language modeling and expert-specific supervision:

$$\begin{aligned} \mathcal{L}_{\mathrm{CoVT}} = \mathcal{L}_{\mathrm{CE}} + \gamma\Big(&\lambda_{\mathrm{seg}}\mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}} \\ &+ \lambda_{\mathrm{edge}}\mathcal{L}_{\mathrm{edge}} + \lambda_{\mathrm{dino}}\mathcal{L}_{\mathrm{dino}}\Big). \end{aligned} \tag{73}$$

Here, γ balances language and visual supervision, while the individual λ coefficients weight the expert objectives.

At inference, the generated visual thoughts become part of the context for answer generation:

$$\hat{y} = f_\theta\left(I, q, Z^{\mathrm{seg}}, Z^{\mathrm{depth}}, Z^{\mathrm{edge}}, Z^{\mathrm{dino}}\right). \tag{74}$$

Dense outputs can optionally be decoded for interpretation, but they are not required to produce the textual answer. CoVT thus transfers multiple perceptual supervision signals into compact intermediate representations. Its visual objectives concern properties of the observed image rather than an explicit sequence of future scene states.

LVR (Li et al., 2026a) trains autoregressive latent states to reconstruct question-relevant visual evidence. Given an image-question pair (I, q), the visual encoder and projector produce

$$V = \mathrm{Enc}_v(I), \qquad V_T = \mathrm{Proj}(V) = \{v_1, \ldots, v_N\}. \tag{75}$$

During supervised training, annotated regions of interest identify the core visual tokens:

$$\begin{aligned} \mathcal{I}_{\mathrm{roi}} &= \{i : p_i \in \mathrm{ROI}(I, q)\}, \\ V_{\mathrm{core}} &= \{v_i : i \in \mathcal{I}_{\mathrm{roi}}\}, \end{aligned} \tag{76}$$

where p_i is the spatial position associated with visual token i.

After generating the latent-start marker, the model reuses its final hidden state as the input embedding for the following latent step. Writing this recurrence in terms of the latent history gives

$$h_t = f_\theta^{\mathrm{hidden}}\left(V_T, q, h_{<t}\right), \qquad t = 1, \ldots, T_v. \tag{77}$$

The resulting states are supervised to reconstruct the ordered core visual tokens:

$$\mathcal{L}_{\mathrm{LVR}} = \frac{1}{T_v} \sum_{t=1}^{T_v} \|h_t - v_t^{\mathrm{core}}\|_2^2 . \tag{78}$$

Once the latent segment ends, the model resumes text generation. The text objective is

$$\mathcal{L}_{\mathrm{NTP}} = -\frac{1}{T_y} \sum_{t=1}^{T_y} \log p_\theta\left(y_t \mid I, q, h_{1:T_v}, y_{<t}\right). \tag{79}$$

The supervised loss combines both terms:

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{NTP}} + \lambda_{\mathrm{LVR}} \mathcal{L}_{\mathrm{LVR}}. \tag{80}$$

LVR subsequently applies $\mathrm{GRPO}_{\mathrm{latent}}$. During rollout, it records the generated latent states. During optimization, these states are replayed at their original positions, while policy ratios are computed only for generated text tokens. Let $C_{i,t}^{\mathrm{replay}}$ denote the causal prefix for text position t, including previously encountered recorded latent states. The importance ratio is

$$r_{i,t}(\theta) = \frac{\pi_\theta\left(y_{i,t} \mid C_{i,t}^{\mathrm{replay}}\right)}{\pi_{\theta_{\mathrm{old}}}\left(y_{i,t} \mid C_{i,t}^{\mathrm{replay}}\right)}. \tag{81}$$

The trajectory reward combines answer accuracy with a formatting reward that checks the latent boundary markers:

$$R_i = R_{\mathrm{acc}}(y_i) + R_{\mathrm{fmt}}(y_i), \qquad \hat{A}_i = \frac{R_i - \mathrm{mean}_j\, R_j}{\mathrm{std}_j\, R_j}. \tag{82}$$

Writing $\mathcal{T}_i$ for the generated text positions, the clipped objective is

$$\begin{aligned} \mathcal{J}_{\mathrm{LVR}} = \mathbb{E}\Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \Big[\min\Big(& r_{i,t}(\theta)\hat{A}_i, \\ & \mathrm{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i \Big) - \beta D_{\mathrm{KL}}^{i,t} \Big] \Bigg], \end{aligned} \tag{83}$$

where $D_{\mathrm{KL}}^{i,t}$ regularizes the current text-token distribution toward the reference policy.

The region annotations provide supervision for learning latent visual reconstruction, but are not required at inference. The reinforcement learning stage evaluates the final textual output and does not assign an explicit action-probability ratio to the continuous latent positions.

Monet (Wang et al., 2026b) learns continuous visual thoughts through a staged distillation process. Its initial warm-up stage uses interleaved image-text reasoning trajectories containing auxiliary images. The subsequent stages transfer the information supplied by these images into latent embeddings, using observation text tokens to identify the visual evidence that should be preserved.

Let $\mathcal{O}$ denote the positions of text tokens describing relevant observations from an auxiliary image I^{aux}. A teacher that observes the auxiliary image produces hidden representations

$$H^*_{\text{obs}} = \left\{ h^{*,l}_{\text{obs},i} : i \in \mathcal{O},\ l = 1, \ldots, L \right\}, \tag{84}$$

where l indexes model layers. The student introduces a latent segment

$$Z_{\text{latent}} = \{z_1, \ldots, z_K\} \tag{85}$$

and aligns its observation-token representations with those of the teacher:

$$\mathcal{L}_{\text{align-obs}} = \frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \sum_{l=1}^{L} \left[1 - \cos\left(h^{*,l}_{\text{obs},i}, \hat{h}^{l}_{\text{obs},i} \right) \right]. \tag{86}$$

In this stage, auxiliary-image embeddings are accessible to the latent segment but hidden from subsequent text tokens. The controlled information flow is

$$I^{\text{aux}} \longrightarrow Z_{\text{latent}} \longrightarrow \mathcal{O}. \tag{87}$$

Latent-only backpropagation routes the observation-alignment gradients through the generated latent embeddings, preventing this objective from being satisfied solely through changes along the subsequent text-processing path. Combined with next-token prediction, the stage objective is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{NTP}} + \alpha \mathcal{L}_{\text{align-obs}}. \tag{88}$$

The resulting latent representations provide targets for a further stage in which the model no longer receives the auxiliary images. Let $h^{*,l}_{\text{latent},k}$ denote a target hidden state at latent position k and layer l. The model learns to reproduce these representations using

$$\mathcal{L}_{\text{align-latent}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=1}^{L} \left[1 - \cos\left(h^{*,l}_{\text{latent},k}, \hat{h}^{l}_{\text{latent},k} \right) \right], \tag{89}$$

with the supervised objective

$$\mathcal{L}_{\text{stage3}} = \mathcal{L}_{\text{NTP}} + \beta_{\text{align}} \mathcal{L}_{\text{align-latent}}. \tag{90}$$

This procedure first learns which auxiliary-image information should be encoded, and then trains the model to generate the corresponding latent representations without that auxiliary input.

Monet further introduces Visual-latent Policy Optimization (VLPO), which incorporates latent positions into policy optimization. Let $h^{\text{old}}_{i,t}$ be a recorded rollout latent and $h^{\theta}_{i,t}$ the current model's predicted latent under the same recorded prefix $C_{i,t}$. VLPO constructs a Gaussian-based surrogate likelihood:

$$\widetilde{\pi}_\theta \left(h^{\text{old}}_{i,t} \mid C_{i,t} \right) \propto \exp\left(-\frac{\left\| h^{\text{old}}_{i,t} - h^{\theta}_{i,t} \right\|_2^2}{2\sigma^2} \right), \tag{91}$$

where σ is a predefined scale. The resulting latent-position surrogate ratio is

$$r^{\text{latent}}_{i,t}(\theta) = \exp\left(-\frac{\left\| h^{\text{old}}_{i,t} - h^{\theta}_{i,t} \right\|_2^2}{2\sigma^2} \right). \tag{92}$$

For an interleaved output o_i, the ratio used in policy optimization depends on the position type:

$$r_{i,t}(\theta) = \begin{cases} \dfrac{\pi_\theta(o_{i,t} \mid C_{i,t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid C_{i,t})}, & t \in \mathcal{T}_i, \\ r^{\text{latent}}_{i,t}(\theta), & t \in \mathcal{Z}_i, \end{cases} \tag{93}$$

where $\mathcal{T}_i$ and $\mathcal{Z}_i$ denote text and latent positions, respectively. The clipped objective includes both types:

$$\begin{aligned} \mathcal{J}_{\text{VLPO}} = \mathbb{E}\Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(& r_{i,t}(\theta) \hat{A}_i, \\ & \text{clip}\left(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon \right) \hat{A}_i \Big) - \beta_{\text{KL}} D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}} \right) \Bigg]. \end{aligned} \tag{94}$$

Table 7: Detailed results on dynamic-centric image understanding benchmarks, including BabyVision, Visual Cognitive Tasks, VTBench, OpenWorldQA, and CausalSpatial.

Models	BabyVision					Visual Cognitive Tasks				VTBench			OpenWorldQA	CausalSpatial
	Fine-grained Discrimination	Visual Tracking	Spatial Perception	Visual Pattern Recognition	Overall	Block towers	Michotte	Astronaut	Help or Hinder	Eyeballing Puzzles	Visual Puzzles	Overall		
Qwen3-VL-8B	7.36	7.23	6.59	9.80	7.47	51.48	19.81	23.62	16.29	13.81	43.00	23.22	32.00	36.96
Cosmos3	7.98	8.43	9.89	7.84	8.51	53.06	19.02	23.49	16.02	16.67	39.00	23.87	35.80	38.44
Think About Image														
+Vision-R1 (ICLR2026)	9.20	8.43	10.99	11.76	9.79	52.30	19.45	25.14	17.28	16.19	40.00	23.87	36.00	38.04
+ViGoRL (NeurIPS2025)	9.82	7.23	8.79	11.76	9.28	53.97	18.38	23.52	17.64	15.24	39.00	22.90	35.20	37.55
Think With Image														
+ControlMLLM (NeurIPS2025)	8.59	10.84	9.89	13.73	10.05	51.27	18.90	24.38	17.40	14.76	42.00	23.55	36.20	38.93
+DyFo (CVPR2025)	9.82	8.43	10.99	9.80	9.79	52.96	19.93	25.16	17.53	15.71	43.00	24.52	37.40	38.83
+CoFFT (NeurIPS2025)	7.98	9.64	13.19	5.88	9.28	53.61	20.85	25.67	16.29	16.19	44.00	25.16	36.00	40.02
+DeepScan (CVPR2026)	7.36	9.64	7.69	13.73	8.76	53.75	20.56	26.10	14.94	15.24	43.00	24.19	36.20	37.85
+FOCUS (NeurIPS2025)	10.43	10.84	8.79	7.84	9.79	54.60	21.87	25.49	16.85	16.67	44.00	25.48	36.40	38.44
+VISER (NeurIPS2025)	9.20	7.23	10.99	11.76	9.54	52.57	21.02	25.78	15.57	18.10	45.00	26.77	37.60	40.32
+ICoT (CVPR2025)	8.59	8.43	9.89	9.80	9.02	53.21	20.34	25.01	17.11	16.67	43.00	25.16	37.00	39.53
Latent Reasoning Method														
+CoVT (ECCV 2026)	10.43	12.05	12.09	7.84	10.82	55.03	21.31	24.30	16.50	17.61	39.00	24.52	36.40	41.01
+LVR (ICLR 2026)	9.20	10.84	8.79	11.76	9.79	55.66	21.29	24.92	17.25	19.52	43.00	27.10	37.00	39.62
+Monet (CVPR 2026)	11.04	12.05	9.89	9.80	10.82	56.19	22.18	24.79	16.55	18.57	41.00	25.81	38.20	41.11
Our Method														
+UnI-R	**17.18**	**16.87**	**18.68**	**15.69**	**17.27**	**62.35**	**26.53**	**29.09**	**22.70**	**22.86**	**49.00**	**31.29**	**41.40**	**43.28**

Advantages are computed from group-normalized rewards combining answer accuracy and output formatting:

$$R(o_i) = R_{\text{acc}}(o_i) + R_{\text{fmt}}(o_i). \tag{95}$$

Unlike an objective restricted to text-token ratios, VLPO also connects the trajectory reward to deviations in continuous latent embeddings through its surrogate ratio. The Gaussian formulation is used for policy optimization and does not imply that auxiliary images are available during final inference.

C.3 Detailed Hyperparameters

For the main comparison, UnI-R and the adapted reasoning baselines use Cosmos3-Nano as the common backbone, with all training and evaluation conducted on eight NVIDIA A100 GPUs. We train UnI-R with CDFL for two epochs, followed by one epoch of CFGSD, using a learning rate of 1×10^{-5} and a weight decay of 0.01. During both stages, we optimize the Foresight-to-Reasoning Bridge and the final autoregressive MLP while keeping the remaining backbone parameters frozen. The default configuration generates $T = 32$ future frames, retains $K = 256$ future hidden states through ACDC, and aggregates them using $R = 32$ learnable queries and $L_R = 3$ readout Transformer blocks. These settings remain unchanged between training and inference unless otherwise specified in the ablation studies. For the CFGSD stage, we follow the standard hyperparameter settings of the reference OPSD implementation for on-policy sampling and distillation. During evaluation, we follow the official settings of all compared methods and use their publicly released code to ensure a fair comparison, where all methods are evaluated on the same Cosmos3-Nano backbone. Training-free methods are applied without additional parameter updates, while trainable baselines follow their respective training procedures. We set `do_sample` $=$ `False` and `num_beams` $= 1$ for autoregressive answer generation, corresponding to greedy decoding and deterministic token selection for fixed inputs and visual conditioning.

D Additional Results

D.1 Detailed Results on Dynamic-Centric Image Understanding

We extend the comparison in Table 1 with additional task-level results. Table 7 presents the full BabyVision category breakdown, four Visual Cognitive Tasks, the two VTBench subsets and their overall score, and results on OpenWorldQA and CausalSpatial. The additional results cover Astronaut, Help or Hinder, Eyeballing Puzzles, and Visual Puzzles, which are not individually reported in the main table. nI-R achieves the highest score across all 14 reported metrics. On BabyVision, it outperforms Cosmos3 in all four ability categories, with improvements ranging from 7.85 to 9.20 points. Its overall score reaches 17.27, exceeding Cosmos3 by 8.76 points and the strongest compet-

Table 8: Number of learnable queries R.

Setting	VTBench	V*	MME-RealWorld	SpatialEval
Cosmos3	23.87	75.92	44.40	64.69
0	24.19	75.92	44.66	65.02
8	29.03	80.10	51.33	67.94
16	30.00	**81.68**	51.17	68.35
24	30.32	80.63	52.21	68.50
32	31.29	81.16	**52.74**	**68.54**
40	**31.94**	79.58	52.16	68.37
48	31.61	78.01	52.27	67.65
56	30.65	78.53	51.95	68.01
64	30.97	78.53	51.54	67.26

Table 9: Number of readout blocks L_R.

Setting	VTBench	V*	MME-RealWorld	SpatialEval
Cosmos3	23.87	75.92	44.40	64.69
1	25.16	78.01	45.44	66.50
2	29.68	80.63	51.43	68.02
3	31.29	**81.16**	**52.74**	68.54
4	**31.93**	**81.16**	52.21	68.66
5	31.61	80.10	51.59	68.82
6	**31.93**	79.06	51.43	68.73
7	30.65	80.10	51.59	**69.20**
8	30.97	80.10	50.81	68.98
9	31.29	79.58	51.07	68.66

Table 10: Number of future frames T.

Setting	VTBench	V*	MME-RealWorld	SpatialEval
Cosmos3	23.87	75.92	44.40	64.69
0	24.19	75.92	44.66	65.02
8	26.45	80.63	47.16	67.35
16	28.81	82.20	51.28	68.08
24	29.68	**82.72**	51.75	67.90
32	31.29	81.16	**52.74**	68.54
40	31.61	79.58	52.11	**68.92**
48	**32.58**	80.10	51.80	68.36
56	31.29	79.58	52.11	67.90
64	31.94	78.53	51.02	66.81

Table 11: Retained token budget K.

Setting	VTBench	V*	MME-RealWorld	SpatialEval
Cosmos3	23.87	75.92	44.40	64.69
0	24.19	75.92	44.66	65.02
8	21.94	76.96	47.06	62.71
16	24.52	76.44	48.20	65.89
32	25.48	78.01	50.29	67.10
64	29.03	81.16	50.70	68.45
128	30.65	**81.67**	51.95	68.27
256	**31.29**	81.16	**52.74**	68.54
512	**31.29**	79.58	52.06	**68.76**
1024	30.97	78.53	52.27	68.60
2048	30.65	79.58	52.16	68.35

Blue shading indicates the default configuration. Bold indicates the highest score in each column within each parameter sweep, including ties.

ing result by 6.45 points. The improvements also extend to OpenWorldQA and CausalSpatial, with gains of 5.60 and 4.84 points over Cosmos3, respectively.

The newly reported tasks further broaden this comparison. On Astronaut and Help or Hinder, UnI-R achieves 29.09 and 22.70, improving over Cosmos3 by 5.60 and 6.68 points and exceeding the strongest competing result on each task by 2.99 and 5.06 points, respectively. On Eyeballing Puzzles and Visual Puzzles, UnI-R obtains 22.86 and 49.00, corresponding to gains of 6.19 and 10.00 points over Cosmos3 and margins of 3.34 and 4.00 points over the strongest competing results. Thus, the VTBench improvement is accompanied by gains on both constituent subsets, while the additional Visual Cognitive Tasks show consistent benefits beyond the tasks reported in the main table.

D.2 Detailed Hyperparameter Ablation

We provide the numerical results underlying Figure 4. Tables 8–11 report the effects of the number of learnable queries R, the number of readout Transformer blocks L_R, the number of generated future frames T, and the retained token budget K, respectively. Each parameter is varied independently while the others remain at their default values: $R = 32$, $L_R = 3$, $T = 32$, and $K = 256$. Cosmos3 is included as a reference in each table. Increasing readout capacity initially improves performance, but the gains are not monotonic. For learnable queries, $R = 32$ achieves the highest scores on MME-RealWorld and SpatialEval, whereas VTBench and V* peak at $R = 40$ and $R = 16$, respectively. Increasing the number of readout blocks from one to three improves all four benchmarks. Further increases yield smaller, benchmark-dependent gains: SpatialEval peaks at seven blocks, while MME-RealWorld peaks at three. These results indicate that effective readout is not restricted to a single query count or depth.

A similar pattern holds for the future information budget. Increasing the number of future frames improves performance over the zero-frame setting, although the preferred count varies across benchmarks. Very small token budgets can limit performance, while increasing K beyond 256 provides no consistent benefit. Overall, these results suggest that UnI-R remains effective across a broad range of evaluated settings, with limited sensitivity to individual hyperparameter choices rather than dependence on a narrowly tuned configuration. The shared defaults achieve the highest MME-RealWorld score in all four sweeps and remain competitive on the other benchmarks, supporting their use as a practical compromise without benchmark-specific hyperparameter tuning.

E Additional Case Studies

Dynamic scenarios. Figure 6 presents five examples involving tower stability, falling objects, collision risk, and ball trajectories. The baseline responses primarily rely on the current scene configuration, interpreting visible overlap, apparent separation, or instantaneous motion as sufficient evidence for predicting subsequent events. In contrast, the UnI-R responses consider how support relationships, contact events, and trajectories may evolve. For example, a stable tower base does not guarantee that the upper blocks remain supported, while apparent separation between objects does not exclude a collision along the direction of travel. Similarly, predicting where falling food first makes contact requires considering its path relative to the pot opening rather than only its initial alignment. The soccer and basketball responses account for the descending arc relative to the target, describing scoring as more likely rather than impossible. Together, these cases illustrate the intended role of foresight in connecting observed spatial configurations with plausible physical outcomes, beyond describing the scene at a single moment.

Static scenarios. Figure 7 presents five examples of fine-grained text recognition and contextual visual grounding, covering advertising displays, store logos, product labels, and signs. The visualized sequences depict progressively closer views of the queried regions, emphasizing details that occupy only a small part of the original image. The baseline responses illustrate two distinct errors: confusing visually similar character sequences and reading a prominent but irrelevant sign. The UnI-R responses instead identify the requested lettering while maintaining its association with the correct visual target. For instance, distinguishing "Panasonic" from "Panasonix" requires character-level precision, whereas identifying "POSTKAR" requires locating the storefront sign specified by the question rather than selecting the more prominent pavement sign. These examples therefore involve both resolving local visual details and interpreting the spatial references that determine which details matter. They illustrate how changes in viewpoint can support closer inspection of question-relevant regions, motivating a role for foresight in static visual understanding as well as physical prediction.

F Limitations and Future Work

Although UnI-R achieves strong performance across diverse visual reasoning tasks, it currently uses a fixed future latent budget, which may not fully match the varying reasoning demands of different questions. More complex questions may benefit from longer future sequences and more retained latent states, whereas simpler questions may require only a limited amount of dynamic information. Adapting the future horizon and latent budget to task difficulty is therefore a promising direction for future work. UnI-R is naturally compatible with such an extension, as its unified latent interface can accommodate varying amounts of future information while preserving the core connection between visual generation and visual reasoning.

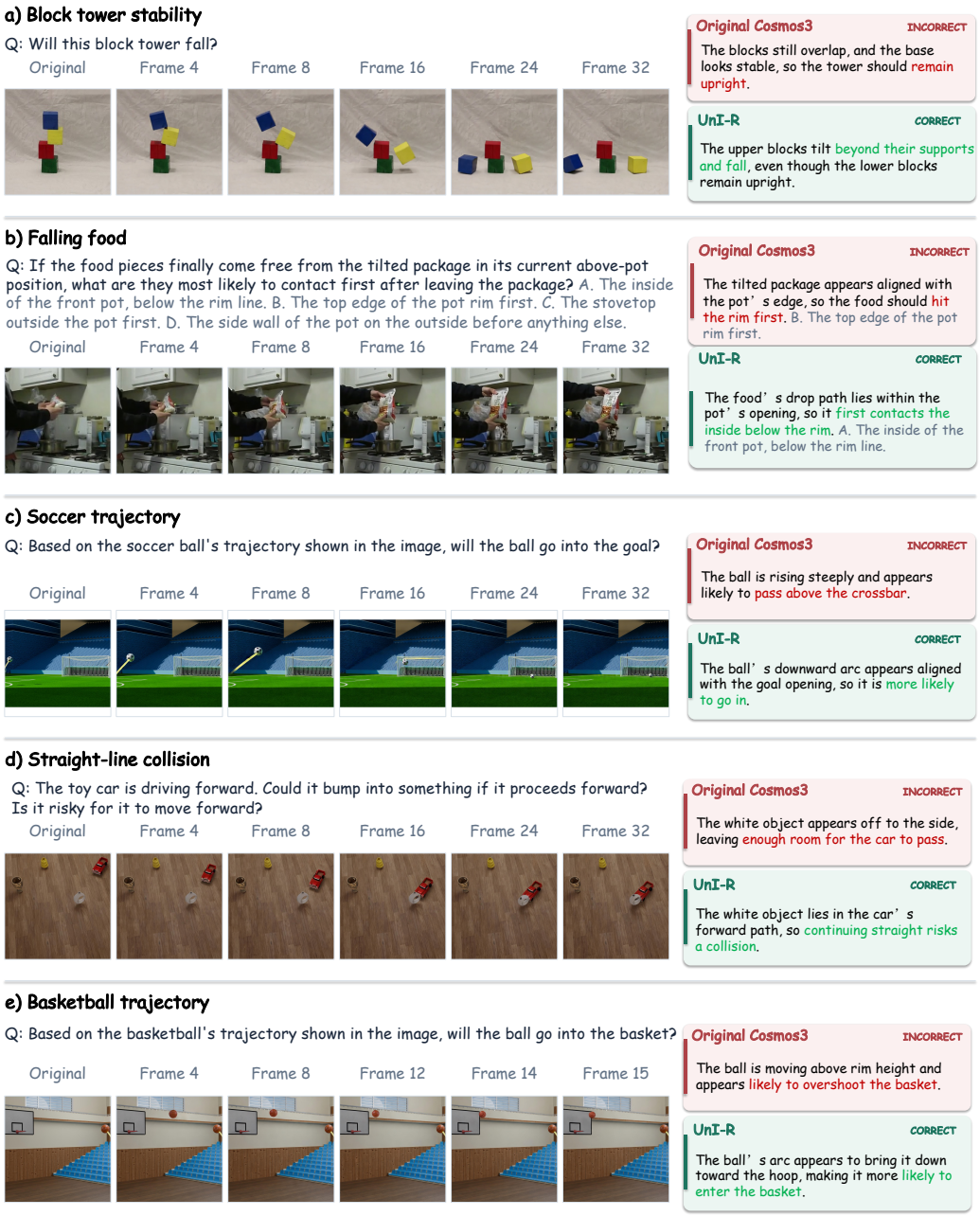


Figure 6: **Additional illustrative cases of UnI-R on dynamic visual questions.** The examples cover (a) block tower stability, (b) falling food, (c) soccer trajectories, (d) straight line collisions, and (e) basketball trajectories. Comparisons with the original Cosmos 3 illustrate how imagined scene evolution can support judgments about stability, motion, and interaction outcomes beyond static observations.

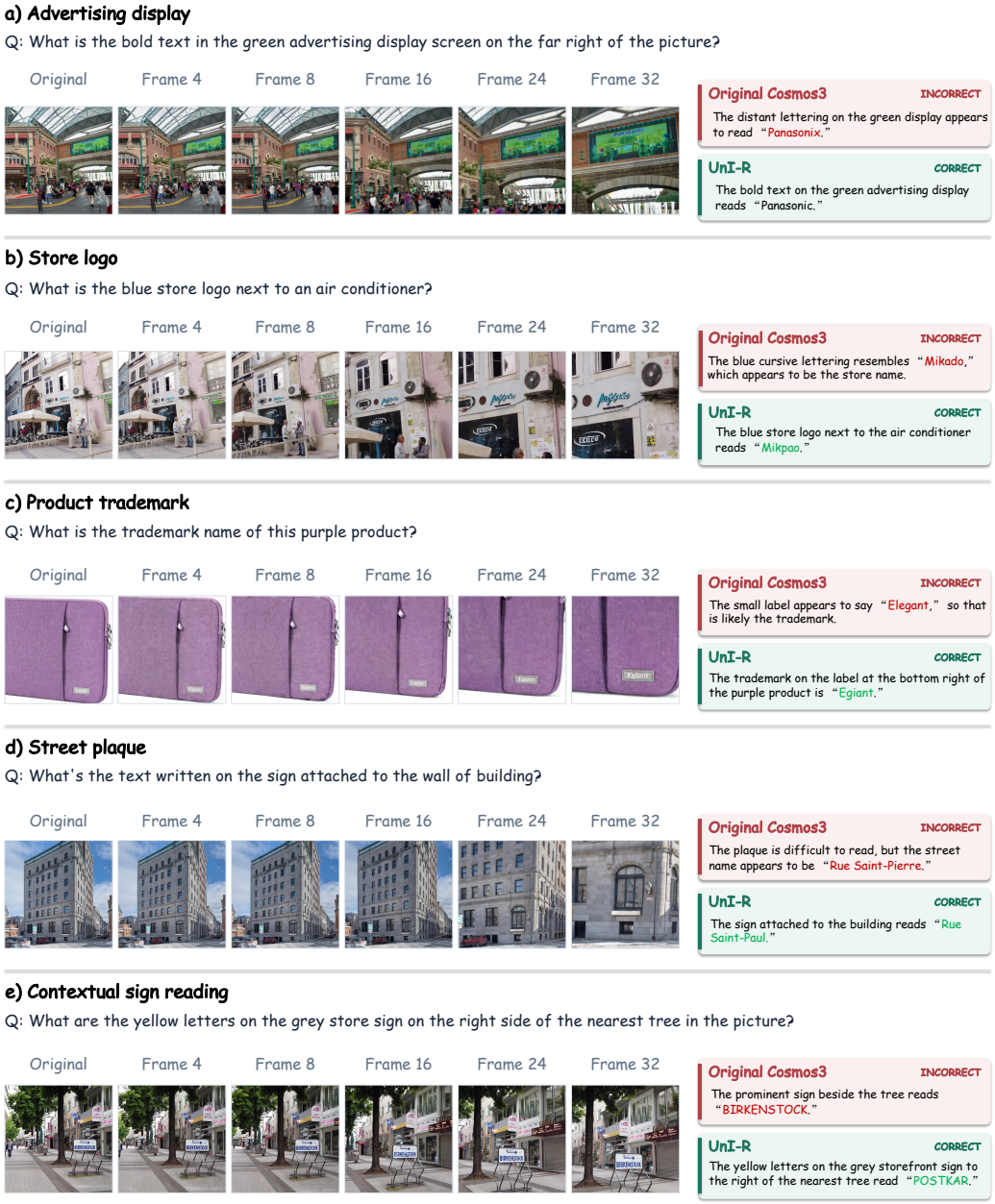


Figure 7: **Additional illustrative cases of UnI-R on general visual questions.** Imagined rollouts progressively focus on relevant regions across (a) advertising displays, (b) store logos, (c) product trademarks, (d) street plaques, and (e) contextual sign reading. Comparisons with the original Cosmos 3 illustrate how latent foresight can support detailed text recognition and the identification of relevant visual evidence.