UniVR: Thinking in Visual Space for Unified Visual Reasoning

Zhongwei Ren[1,2,*] **Yunchao Wei**[1,3] **Yao Zhao**[1] **Weibo Gong**[2]
Xiao Liu[2] **Anran Wang**[2] **Xiangtai Li**[2,*,†] **Xiaojie Jin**[1,*]
[1]Beijing Jiaotong University [2]ByteDance
[3]Beijing Academy of Artificial Intelligence, BAAI
[*]Equal contribution [†]Project lead
Correspondence: Xiaojie Jin, Yunchao Wei `https://UniVR.github.io/`

Abstract

Learning broad world knowledge directly from raw visual data is a fundamental capability of intelligence. We introduce **UniVR**, which investigates simultaneously learning complex reasoning, fine-grained physical dynamics, and long-term planning from text-instructed visual demonstrations. At its core, UniVR features VR-GRPO, a reinforcement learning paradigm with complementary global and step-level rewards. This approach promotes logical coherence and physical consistency through VLM-mediated feedback, without using explicit textual chains of thought as intermediate reasoning traces. To train and evaluate UniVR, we construct **VR-X**, a large-scale benchmark curated from 16 diverse sources spanning long-horizon manipulation, spatial puzzles, and physical reasoning. It assesses these heterogeneous capabilities through generated visual trajectories. UniVR improves the overall VR-X score by 18.4 points over its Emu3.5 backbone, and its enhanced visual reasoning also benefits multimodal understanding benchmarks. These findings underscore the potential of reasoning within visual spaces. Code, data, and models are planned for public release.

1 Introduction

Current AI models primarily derive world knowledge from text [39, 6, 64, 73, 74, 2, 4], performing reasoning [29, 27, 35, 57] and planning [32, 66, 68, 84] within the textual space. However, text is an abstract representation of the world, which is unable to fully encompass the rich information of the real visual world, such as complex dynamics, spatial relationships, and underlying physical laws. In contrast, vision serves as the most direct medium for world knowledge and remains the primary source for animals and humans to acquire information. In most scenarios, humans can perform complex reasoning without relying on language by directly simulating task execution and scene transitions in their minds, which constitutes our innate visual reasoning capability. Given the vast abundance of video content available on the internet, equipping AI with the capacity for complex reasoning and planning within the visual space holds significant promise for enhancing its world modeling abilities and pushing the frontier of efficient task execution in the real visual world.

Recent research has explored two primary pathways for advancing visual reasoning. On one hand, MLLMs, such as Gemini-3 [23] and GPT-5 [67], can generate detailed textual reasoning chains that are subsequently converted into visual reasoning traces via generative models like Nano Banana [14] and GPT-Image [50]. While this paradigm conveniently leverages pre-existing linguistic strengths, textual abstractions inherently struggle to capture the intricate dynamics and spatial relationships of the physical world. Such limitations prevent a seamless transfer of textual reasoning proficiency into the visual domain. As shown in Fig 1, even with state-of-the-art reasoning and high-fidelity rendering, models frequently fail to maintain logical coherence and physical consistency in tasks requiring

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

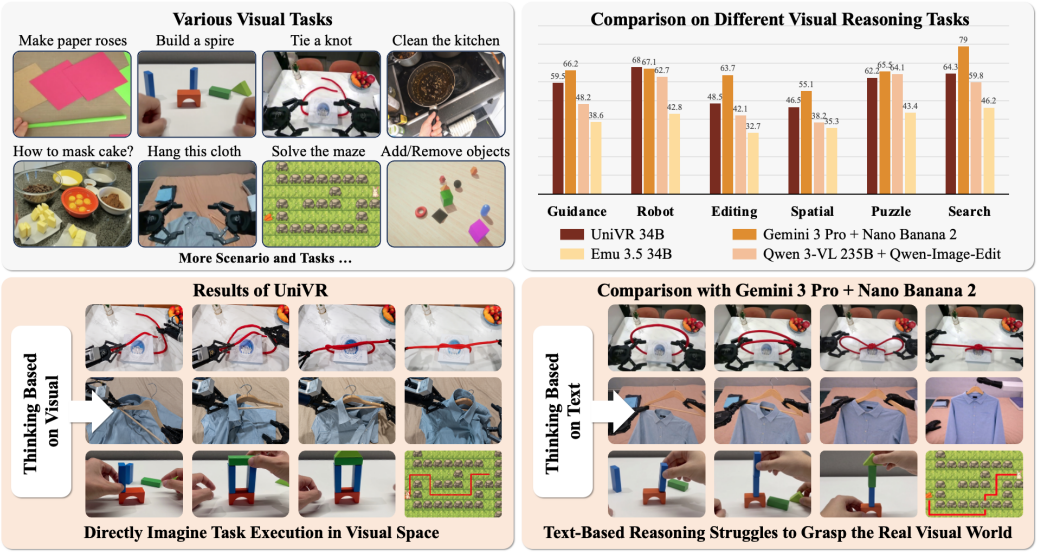

Figure 1: UniVR performs native visual-space reasoning and planning across diverse tasks, improving on Emu3.5 and complementing text-mediated reasoning pipelines.

fine-grained, long-horizon visual evolution. On the other hand, some works [59, 58, 70, 49, 60] leverage video generation models to model task strategies directly in visual space, while others employ unified generation models [15, 81, 90, 17, 87] for more comprehensive reasoning through joint vision-language representations. Despite their promise in acquiring complex knowledge from visual data, their reasoning still remains heavily dependent on text guidance, requiring dense image-text pairs for both training and inference. These limitations constrain the scaling of visual reasoning and drive our core inquiry: **How can models utilize raw visual data to boost their visual reasoning capabilities across diverse tasks?**

To investigate this problem, we first establish VR-X, a comprehensive benchmark designed to assess visual reasoning capabilities across diverse tasks. As shown in Fig 4, VR-X encompasses two primary task categories. The first focuses on long-horizon complex planning across various environments, featuring minute-scale tasks with fine-grained dynamics (*e.g.* tying knots, folding clothes) and diverse domains spanning cooking, crafts, robotic control, and navigation. The second focuses on general visual reasoning, evaluating fundamental cognitive skills such as visual search, puzzles, spatial perception, and editing. Models are tasked to directly reason and generate in visual space, with assessment centered on the logical coherence and task completion of the visual reasoning trace. Such task diversity and evaluation paradigm is rarely explored in existing benchmarks, demanding robust learning capabilities to master heterogeneous task knowledge.

Based on this benchmark, we evaluate state-of-the-art MLLMs, including Gemini [23, 14], GPT [67], and Qwen [54, 3], by prompting them to generate detailed textual reasoning chains to guide generative models, alongside visual generation models such as Emu3.5 [15]. Results indicate that while they excel at textual reasoning, visual comprehension, or high-fidelity generation, they still struggle to accurately execute tasks in the benchmark. As shown in Fig. 1 and Fig. C.4, their predictions contain errors such as logical gaps, physical inconsistency, and violations of rules, ultimately failing to produce visually coherent task sequences. This underscores an urgent need for frameworks that effectively master those heterogeneous knowledge.

Motivated by these observations, we propose **UniVR**, a unified next-token prediction framework with strong visual reasoning capabilities. At its core is VR-GRPO, a novel RL paradigm that learns diverse knowledge directly in visual space. A VLM evaluator first provides holistic assessment of generated sequences for task completion and visual quality. However, as detailed in Sec. 3.2, this vanilla reward fails to capture errors like logical gaps and physical inconsistency. We attribute this to current VLMs' limited ability to identify fine-grained errors in minute-level multi-step sequences, leaving final states and appearance quality to dominate judgments. We therefore propose a **Step-Focal** reward that targets uncertain substeps for more precise assessment. Combined with global evaluation, this design promotes both overall task completion and fine-grained reasoning coherence without task-specific

reward rules. The pipeline still uses textual instructions and VLM-based supervision; visual reasoning refers to its intermediate representation, not the absence of language from training.

In addition to these qualitative analyses, we benchmark UniVR against existing methods on **VR-X**. UniVR improves on its Emu3.5 backbone and remains competitive with strong text-mediated pipelines on long-horizon manipulation tasks. These comparisons do not isolate reasoning modality: the models differ in capacity, training data, and rendering modules. We therefore also compare unified models fine-tuned on the same VR-X data and validate the findings using independent evaluator families and repeated RL runs (Sec. 5.1). Further visualizations are provided in Fig. C.3.

Our contributions are summarized as follows:

- We explore learning heterogeneous tasks, from long-term planning to general cognitive reasoning, through visual state transitions rather than explicit textual chains of thought.
- We propose VR-GRPO, combining a Step-Focal reward on uncertain reasoning substeps with a global reward. VLM-mediated process feedback improves logical coherence and physical consistency without task-specific reward rules.
- We introduce VR-X, a benchmark for evaluating diverse tasks in visual space, spanning from fine-grained long-term planning to general reasoning, to facilitate future research on visual reasoning.

2 Related Works

Visual Reasoning. The rapid advancement of LLMs [13, 6, 2, 93, 33, 79, 94] and MLLMs [4, 88, 40, 3, 72] has spurred extensive research on text-centric reasoning paradigms, such as chain-of-thought [102, 85, 48] and latent-space reasoning [28, 12]. Recent efforts like visual CoT [37, 80, 53] extend this paradigm to multimodal inputs, yet they still project visual features into a linguistic space, leaving reasoning fundamentally text-bounded . Another line of work explores video generation for non-linguistic world modeling in autonomous driving [49, 60] and robotics [70, 26]. However, these approaches are typically confined to short-horizon dynamics or narrow, single-task settings. UniVR departs from both paradigms by directly reasoning in the visual space, unifying and improving heterogeneous tasks, which demand long-horizon planning, complex policy and intricate physics.

Unified Model [38, 17, 31, 90, 87, 52, 96] aims to encode world knowledge in a single model space, combining the textual reasoning of MLLMs with the visual generation capabilities. Pioneering architectures [15, 81] tokenize images, text, and video into a unified discrete space to enable flexible cross-modal generation. However, these models are predominantly trained on entertainment-oriented or artistic editing objectives, and their visual knowledge acquisition remains heavily constrained by dense image-text supervision. Our UniVR breaks this dependency by incorporating a text-free visual reasoning framework into the unified architecture, enabling the model to learn complex reasoning and planning directly from raw visual inputs.

Reinforcement Learning in Generative Model. The success of RL in MLLMs [61, 56, 51, 19, 97, 39, 98, 71, 43] has been extended to visual generation, with pioneering works such as DDPO [5] and ReFL [91] employing PPO [61] or RLHF [51] to align models with human preferences regarding image fidelity. Following the success of DeepSeek-R1 [25], GRPO [64, 41] has been widely adopted across various vision tasks to enhance multimodal understanding and image/video generation. Various approaches [83, 92, 100] also incorporate specialized reward, such as boosting aesthetic appeal via HPSv3 [47, 89] or enforcing semantic alignment through CLIP [55] and VLM-based scoring. However, these reward mechanisms primarily target perceptual quality, text-image consistency, or single-step correctness. In contrast, our VR-GRPO is specifically designed to optimize for logical consistency and physical plausibility in the context of multi-step visual reasoning and policy learning.

3 UniVR

UniVR (Fig. 2) uses autoregressive visual reasoning with cold initialization (Sec. 3.1) followed by VR-GRPO (Sec. 3.2).

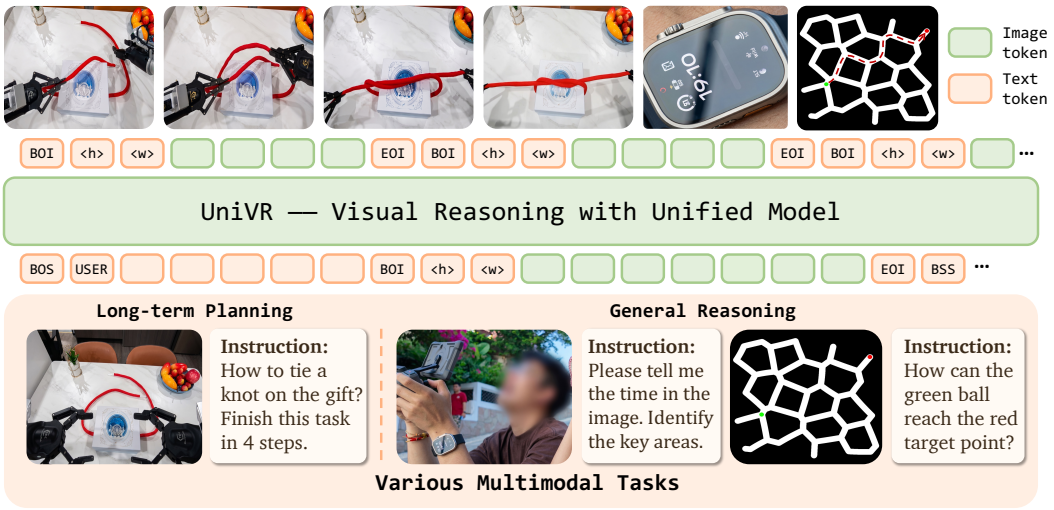

Figure 2: **UniVR architecture.** Unified next-token prediction generates visual reasoning traces from instructions and query images.

3.1 Thinking in Visual Space

Given an image sequence $x_{1:t}$ and an instruction, we model the next-frame distribution: $p(x_{t+1} \mid x_{1:t})$. We use image sequences containing demonstration trajectories of task execution across diverse scenarios, encompassing various kinds of planning and reasoning knowledge. Our formulation does not require dense textual reasoning chains, but instead directly models the state transitions and underlying policy dynamics in these trajectories, encouraging the model to reason in visual space.

Specifically, we adopt Emu3.5 as our baseline, a state-of-the-art unified generative model that produces variable-length image sequences. It employs a VQ-VAE-style autoencoder [77] to encode images and text into a unified discrete vocabulary, also enabling text generation. This allows us to further investigate how enhanced visual reasoning affects multimodal understanding (See Sec. 5.2). Despite pretraining on extensive image-text pairs, this baseline still struggles to capture complex real-world physics, fine-grained action dynamics, and visual cognitive abilities, as shown in Sec. 5.1 and Fig. C.4. We attribute this to its heavy reliance on dense textual reasoning chains for world knowledge acquisition, reducing vision to a mere renderer. We therefore first perform supervised fine-tuning on a curated dataset of diverse visual tasks as cold initialization, standardizing all samples as query image, instruction, visual reasoning trajectory. This endows the model with visual reasoning priors for subsequent RL. Training configuration and data construction are detailed in Appendix A.

3.2 Visual Reasoning GRPO

Despite the versatility of casting various tasks into a unified visual space, we observe that vanilla SFT still struggles to reconcile heterogeneous multi-source tasks. For instance, 2D puzzles and long-term human planning ediffer drastically in temporal scale, domain knowledge, and visual appearance. As detailed in Sec. 5.2, dense supervision in SFT training fails to yield consistent improvements across diverse tasks. While RL method, *e.g.* GRPO [64], offer a promising paradigm for performance enhancement by enabling models to autonomously explore optimal policies across diverse scenarios, existing methods prioritize visual fidelity or cross-modal alignment rather than visual reasoning.

To address this, we introduce VR-GRPO, an RL methodology that optimizes visual reasoning trajectories using textual instructions and VLM-mediated rewards, rather than explicit textual reasoning traces or task-specific reward rules. In Sec. 5.2, we further demonstrate that VR-GRPO can also seamlessly integrate with textual reasoning tasks, facilitating a synergistic advancement in overall multimodal understanding.

Reward design. The VR-GRPO has two reward components: format reward R_{format} and visual reasoning reward R_{reason}. R_{reason} consists of a global reward R_{g} and a novel step-focal reward R_{s}. R_{format} ensures that the generated image sequences satisfy structural constraints, such as uniform resolution and the number of reasoning steps prescribed by the task instructions.

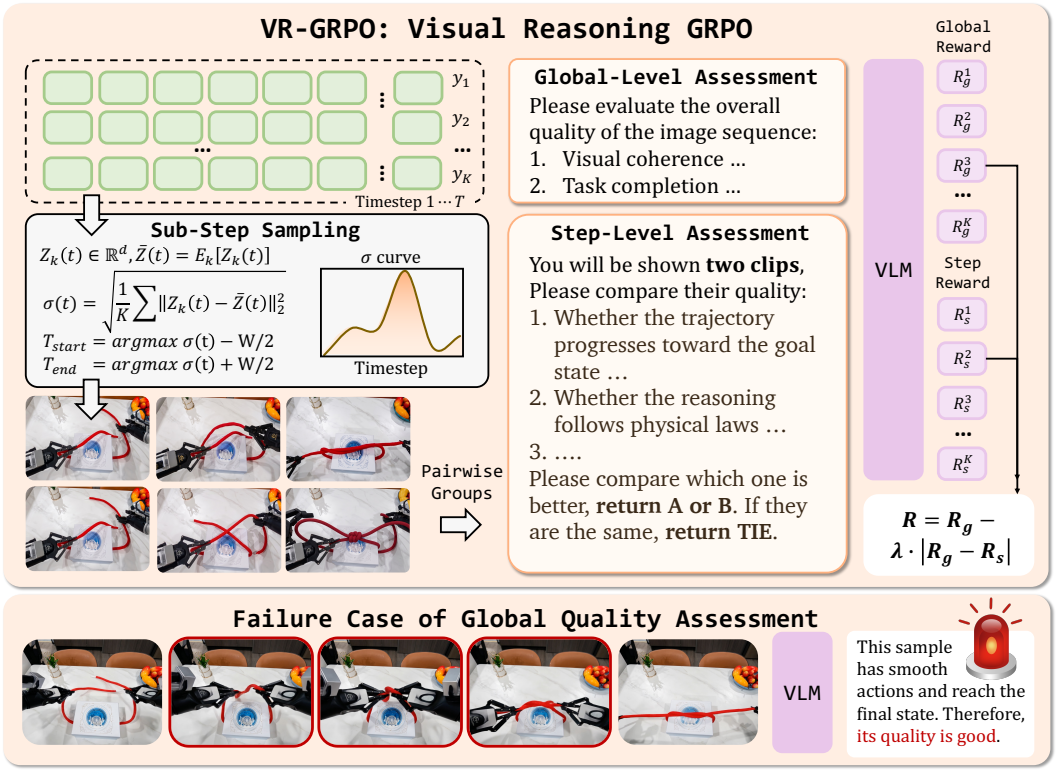

Figure 3: **Visual Reasoning GRPO.** Global and Step-Focal feedback assess task completion and physical coherence (top); global-only judgments can overlook local errors (bottom).

For R_{reason}, we use a general-purpose prompt to guide an online reward model (*i.e.* Qwen3-VL-30B) in providing an overall quality assessment of rollout samples against ground-truth references. While the VLM effectively assesses task completion and visual fidelity, as shown in Fig. C.2, our preliminary analysis indicates that this global reward often overlooks intermediate physical violations and logical gaps, over-prioritizing terminal success and pixel clarity, especially in long-horizon, multi-step tasks. We attribute this to current VLMs' reliance on text-derived world knowledge, which constrains their ability to pinpoint fine-grained visual dynamic errors in minute-level multi-step sequences.

Consequently, we introduce a step-focal reasoning reward R_s that prioritizes uncertain steps in the reasoning trajectory. By focusing the VLM on these steps, we provide localized feedback on logical and physical coherence throughout long-horizon tasks. Specifically, for a set of K rollout trajectories $y_{1:K}$, we first assume a uniform length T. We employ a CLIP image encoder to extract per-frame feature embeddings as $z_k(t) \in \mathbb{R}^d$ and calculate the inter-trajectory variance at each timestep t as:

$$\sigma(t) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|\mathbf{z}_k(t) - \bar{\mathbf{z}}(t)\|_2^2} \tag{1}$$

where $\bar{\mathbf{z}}(t)$ is the mean embedding across all trajectories at time t. A high variance $\sigma(t)$ indicates a state of maximum uncertainty where the model's reasoning paths diverge. Given a window size W, we identify the peak uncertainty at $t^* = \text{argmax}_t \sigma(t)$ and extract a reasoning segment from $[t^* - W/2, t^* + W/2]$. If uncertainty is nearly uniform across time, peak selection becomes uninformative; this is a limitation of the signal, not a separate task-specific random-sampling rule. In practice, generated trajectories often vary in length. To ensure temporal alignment, we partition each trajectory into an equal number of segments. We then treat the segment index as the synchronized timestep and use the average CLIP features within each segment to perform the uncertainty analysis described above. Following [82], we employ a pairwise evaluation protocol for both global rewards and the identified sub-steps to derive relative win rates. For either comparison scope $a \in \{g, s\}$, the

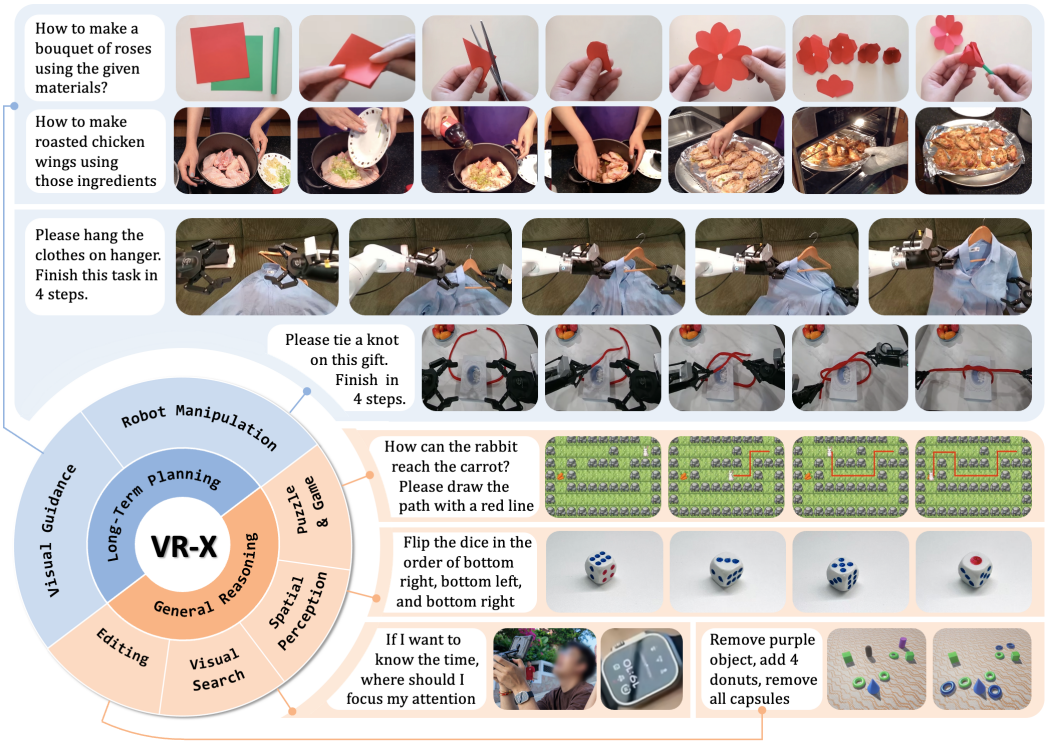

Figure 4: **Overview of VR-X benchmark.**

evaluator returns Win, Tie, or Loss, and

$$R_a(y_k) = \frac{1}{K-1} \sum_{j \neq k} \left[\mathbb{1}\{E_a(y_k, y_j) = \text{Win}\} + \frac{1}{2} \mathbb{1}\{E_a(y_k, y_j) = \text{Tie}\} \right]. \tag{2}$$

Thus both rewards lie in $[0, 1]$, and ties need not force a preference between equally valid strategies. Appendix A.3 specifies temporal alignment, aggregation, and the evaluation prompt. Upon obtaining the global score R_{g} and step-focal scores R_{s} for each trajectory, we integrate them using the following formula:

$$R_{\text{reason}} = R_{\text{g}} - \lambda |R_{\text{g}} - R_{\text{s}}| \tag{3}$$

The coefficient λ penalizes disagreement between global completion and local process assessments, discouraging terminal-state shortcuts.

4 VR-X Benchmark

Our objective is to advance the capacity for reasoning natively within the visual domain across a wide array of scenarios. We prioritize a comprehensive benchmark capable of evaluating multi-step planning, delicate manipulations and cognitive reasoning. However, current benchmarks mostly emphasize visual quality or text-matching, limited to restricted environments and short-term tasks.

Therefore, we introduce VR-X, the first large-scale benchmark designed for diverse and heterogeneous visual reasoning. As shown in Fig. 4, it includes six tasks, each providing detailed visual reasoning traces: visual guidance, robotic manipulation, puzzle games, editing, search, and spatial perception.

4.1 Dataset Generation

VR-X is characterized by its vast diversity and involves fine-grained, visually complex manipulations that are hard to describe linguistically. We curate 1.5M raw samples from 16 diverse sources (e.g., AgiBot [7], Action100M [10], EgoDex [30], VisualCoT [63]), spanning minute-long planning (robotic manipulation, cooking, handcrafting) to single-step reasoning (mazes, visual search). Rigorously

Table 1: **Comparison on VR-X.** Baselines are zero-shot; UniVR uses VR-X SFT and RL. LMM–T2I pairings are specified in Sec. 5.1. Scores are VLM assessments, not success rates.

Method	Visual Thinking	Long-term planning		General Reasoning				Overall↑	JEPA↓
		Guidance	Robot	Editing	Spatial	Puzzle	Search		
Large Multimodal Model + T2I Model									
Qwen3-VL-235B [3]	✗	48.2	62.8	42.1	38.2	64.1	59.8	52.5	18.08
Qwen3.5-397B [54]	✗	47.0	63.2	39.8	40.4	65.6	64.5	53.4	18.64
GPT-5 [67]	✗	**68.2**	64.1	58.0	49.3	64.0	77.4	63.5	12.17
Gemini-2.5-pro [14]	✗	58.4	**67.9**	54.0	40.5	**67.7**	76.3	60.8	14.39
Gemini-3-pro [23]	✗	66.2	67.1	**63.7**	**55.1**	65.5	**79.0**	**66.1**	**11.07**
Unified Generation Model									
Janus-pro [11]	✗	9.2	18.2	5.4	10.2	27.1	21.5	15.3	68.79
Show-o2 [90]	✗	15.1	22.5	13.0	17.1	29.4	35.8	22.2	59.93
ILLUME+ [31]	✗	13.1	11.5	5.8	14.6	22.2	27.5	15.8	61.12
OneCAT [38]	✗	15.6	13.5	10.3	20.2	16.2	30.1	17.7	77.06
STAR [52]	✗	22.7	28.5	14.2	21.5	27.2	37.7	25.3	51.98
OmniGen2 [87]	✗	20.4	29.4	15.2	16.9	30.5	42.1	25.8	47.09
Bagel [17]	✗	25.2	34.7	20.9	21.3	35.1	47.7	30.8	40.88
Emu3.5 [15]	✗	38.6	42.8	32.7	35.3	43.4	46.2	39.8	33.62
UniVR	✓	**59.5**	**68.0**	**48.5**	**46.5**	**62.2**	**64.3**	**58.2**	**13.01**
△ *v.s.* Emu 3.5		↑20.9	↑25.2	↑15.8	↑11.2	↑18.8	↑18.1	↑18.4	↓20.61

curated into 310k cold-start training, 3k RL, and 1.8k benchmark evaluation samples, all follow a unified format: query image, textual instruction, visual reasoning trajectory. We further annotate these sequences with fine-grained textual chain-of-thought (CoT) descriptions to support multimodal learning. See Appendix A for more details.

4.2 Evaluation Metrics

Our evaluation focuses on two main aspects: the logical accuracy of visual reasoning and the adherence to real physical dynamics. Accordingly, we employ the following metrics for evaluation:

VLM score: We conduct an automated evaluation using Qwen3.5-397B [54] to quantitatively assess reasoning traces. The model evaluates each sample based on task completion, procedural coherence, visual informativeness, and image fidelity. These dimensions are aggregated into a final normalized score (0–100), providing a robust metric for measuring the logical and visual quality of the results.

JEPA similarity: Despite strong reasoning capabilities, VLMs' dependency on textual knowledge often hinders their ability to perceive intrinsic physical laws. As shown in Fig. C.2, this may result in hallucinations where the evaluator ignores fine-grained physical inconsistencies. To bridge this gap, we incorporate JEPA similarity as an additional metric. Building on prior research [45] demonstrating that V-JEPA [1] encoders can capture high-level latent physical dynamics, we map sequences into latent space and compute maximum mean discrepancy against ground-truth distributions. Lower scores indicate closer alignment with real-world physics. See Appendix A for more evaluation details.

5 Experiment

We leverage Emu3.5 34B to initialize our unified model, which undergoes full-parameter SFT and RL. Our RL pipeline is implemented using the verl framework [65], utilizing a rollout size of 8. For sub-step selection, we set the default window size to 4 frames and the consistency coefficient λ to 2.0. Training resolution, sequence caps, and other hyperparameters are given in Appendix A.

5.1 Results on VR-X

Tab. 1 compares LMM–T2I pipelines and unified generation models on VR-X.

LMMs with T2I models. We first evaluate off-the-shelf LMMs (rows 1-5), including Qwen 3-VL [3] and Qwen 3.5 [54] with Qwen-Image-Edit [86], Gemini 2.5/3 Pro [23, 14] with Nano Banana

Table 2: **Ablation studies.** "LP" and "GR" denote long-term planning and general reasoning in VR-X, respectively.

Method	MMMU [101]	MME(P) [20]	MME(C) [20]	MMBench [42]	MathVista [44]	MM-Vet [99]
Emu 3.5	0.292	781.1	324.6	0.183	41.7	28.0
Text-only training	0.290	782.0	323.4	**0.199**	40.8	28.3
UniVR*	**0.337**	**799.3**	**338.5**	0.198	**44.0**	**35.6**
△*v.s.* Emu3.5	↑ 0.045	↑ 18.2	↑ 13.9	↑ 0.015	↑ 2.3	↑ 7.6

(a) Results on understanding benchmarks. * means training with interleaved image-text sequences.

Cold Start	Reward: Global	Reward: Step	Reward: Pairwise	VR-X: LP	VR-X: GR	VR-X: JEPA↓
✓				48.2	42.4	18.44
✓	✓			45.7	46.0	22.30
✓		✓		54.9	45.8	15.87
✓	✓	✓		61.6	53.7	**12.89**
✓	✓	✓	✓	**63.8**	**55.4**	13.01

(b) Reward in VR-GRPO.

Training Data	Cold Start	Reward: HPSv3	Reward: CLIP	Reward: VR-GRPO	VR-X: LP	VR-X: GR	VR-X: JEPA↓
V	✓				48.2	42.4	18.44
V&T	✓				49.7	41.0	18.37
V&T	✓	✓			49.0	41.7	18.09
V&T	✓	✓	✓		52.6	42.0	17.17
V&T	✓	✓	✓	✓	**65.4**	**57.8**	**12.97**

(c) Training with text-based RL method.

Method	Training Data	LP	GR
Cold-Start	Separate	53.6	44.2
	Join	48.2	42.4
UniVR	Separate	61.5	55.0
	Join	63.8	55.4

(d) Ablation on joint training.

Method	WorldArena [62]	Uni-MMMU [103]	RBench [18]
Emu3.5	40.3	28.7	31.2
Cold-Start	42.4	37.6	36.8
UniVR	49.5	54.4	47.7
△*v.s.* Emu3.5	↑ 9.2	↑ 25.7	↑ 16.5

(e) Evaluation on other benchmark data.

Table 3: **Distribution-matched comparison.** Baselines use the same 310k VR-X samples for SFT; UniVR adds VR-GRPO. LP/GR: long-term planning/general reasoning.

Method	Training	LP↑	GR↑	JEPA↓
Show-o2	VR-X SFT	22.2	28.9	45.92
OmniGen2	VR-X SFT	28.7	33.3	39.56
Bagel	VR-X SFT	46.6	41.3	21.18
Emu3.5	VR-X SFT	48.2	42.4	18.44
UniVR	SFT+VR-GRPO	**63.8**	**55.4**	**13.01**

1/2 [21, 22], and GPT-5 [67] with GPT-image 1.5 [50]. These systems follow a two-stage pipeline: the LMM first produces step-level textual instructions from the input image and task prompt. The generation module then renders the sequence frame-by-frame conditioned on previous frames and textual guidance. Gemini 3 Pro with Nano Banana 2 achieves the best performance, benefiting from strong textual reasoning and superior rendering quality. However, as shown in Fig. C.4, visual inconsistencies persist in tasks with complex dynamics and planning, even with detailed instructions. These qualitative failures motivate visual process supervision, but do not establish that visual reasoning universally outperforms textual reasoning.

Unified generation models. Tab. 1 also evaluates vision-language unified models. Except for Emu3.5, most lack contiguous image generation and must iteratively unroll for long-horizon planning. Emu3.5 achieves the highest overall score in this group (39.8), benefiting from native pre-training on diverse daily-life and handcrafting sequences. Their JEPA scores also lag behind text-mediated pipelines, indicating limitations in modeling visual transitions.

UniVR. UniVR improves the overall score from 39.8 to 58.2, reaching 63.8 on long-term planning and 55.4 on general reasoning. Its JEPA distance also decreases from 33.62 to 13.01. These gains concern visual trajectory quality under the reported protocol, rather than a measured task success rate.

Tab. 3 adapts four unified baselines to the same training distribution. UniVR retains gains over every adapted baseline, including its own SFT initialization. This separates the additional benefit of VR-GRPO from benchmark adaptation, although it does not constitute a compute-matched comparison across architectures. Independent GLM-5 and GPT-5.5 evaluators preserve the method ranking, scoring UniVR at 59.0 and 57.5, respectively, versus 58.2 from Qwen3.5. Their Spearman correlations with blind human judgments are 0.84 and 0.89, versus 0.85 for Qwen3.5 (Appendix B.1). Three

independent RL runs yield an overall score of 58.6 ± 1.3 and JEPA distance of 13.03 ± 0.11 (mean $\pm$ standard deviation across runs; Appendix C.1). Tab. 1 retains the original run for comparability; these standard deviations are not confidence intervals.

5.2 Ablation Study

Visual reasoning benefits multimodal understanding. Tab. 2a compares Emu3.5, text-only training, and UniVR on six multimodal metrics. Following [46], each model generates intermediate visual reasoning images and a final textual answer. Training only on VR-X textual reasoning chains (Appendix A) yields negligible gains. In contrast, UniVR trained with interleaved image–text sequences improves on Emu3.5 across all six metrics, indicating that visual reasoning complements textual supervision.

Reward in VR-GRPO. Tab. 2b ablates the individual components of VR-GRPO. Row one establishes the baseline performance of the cold-start model before RL. While the global reward alone improves general reasoning (short sequences) by 3.6 points, it causes degradation in long-term planning and JEPA scores, indicating reward hacking. Conversely, introducing the step-focal reward alone improves all three metrics. Combining both rewards further improves performance. Finally, transitioning the VLM from absolute scalar scoring to a pairwise comparison protocol improves the final results.

Human analysis provides complementary evidence: uncertainty-peak regions cover 77.3% of annotated errors early in training and approximately 65% late in training. Step-level preferences correlate with human judgments at $\rho = 0.89$, compared with $\rho = 0.76$ for global rankings. These are diagnostic associations, not a guarantee that high variance identifies an error (Appendix C.2).

VR-GRPO is compatible with text-based RL. Tab. 2c investigates the compatibility of VR-GRPO with existing multimodal rewards. In the cold-start phase, adding textual reasoning chains (row 2) yields no significant gains over pure visual reasoning (row 1). Furthermore, incorporating aesthetic-centric rewards like HPSv3 or vision-language alignment via CLIP (rows 3-4) provides limited assistance for tasks requiring complex visual reasoning, with HPSv3 showing negligible impact. In contrast, VR-GRPO (row 5) significantly boosts visual reasoning capabilities and remains compatible with other reward functions, highlighting the generalizability and versatility of our approach.

VR-GRPO stabilizes training with heterogeneous knowledge. Long-term planning and general reasoning in VR-X exhibit significant disparities in terms of temporal scale, environmental context, visual appearance, and knowledge domains. Rows 1-3 in Tab. 2d investigate the joint learning capabilities across these tasks without the incorporation of VR-GRPO. The results indicate that during the pure cold-start phase, joint training struggles to effectively balance heterogeneous task knowledge, failing to surpass the performance of models trained individually on each task. In contrast, the introduction of VR-GRPO in row 4 empowers the model to autonomously explore optimal policies, yielding performance gains in both categories. This demonstrates the efficacy of our approach in facilitating unified visual reasoning training.

Evaluate on other visual reasoning benchmarks. To assess generalization, we evaluate UniVR on three external benchmarks (Tab. 2e): two focusing on embodied intelligence [62, 18] and one on cognitive reasoning [103]. We generate visual trajectories from their instructions and query images and apply the same VLM scoring protocol. Across these different tasks, environments, and viewpoints, UniVR improves on Emu3.5 by up to 25.7 score points and consistently exceeds the cold-start checkpoint, supporting transfer beyond VR-X.

6 Conclusion

In this work, we presented UniVR, a unified framework for complex reasoning and planning through visual state transitions rather than explicit textual reasoning traces. Central to our approach is VR-GRPO, which promotes global task completion and step-level physical coherence through complementary rewards. Trained on text-instructed visual demonstrations and optimized with VLM-based feedback, UniVR improves on unified generation baselines, including those adapted to the same VR-X data. Independent evaluators and repeated RL runs further support these findings, without establishing universal superiority over text-mediated reasoning. Beyond task execution, improved visual reasoning also benefits multimodal understanding, suggesting vision-native computation as a complement to textual abstraction.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 92470203), the Beijing Major Science and Technology Project under Contract no. Z251100008125015 and Beijing Academy of Artificial Intelligence.

References

[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[2] 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*, 2023.

[3] 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.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[9] Brandon Castellano and PySceneDetect Contributors. Pyscenedetect: Video scene detection and analysis tool. GitHub repository, 2025. URL `https://github.com/Breakthrough/PySceneDetect`. Accessed: 2026-05-06.

[10] Delong Chen, Tejaswi Kasarla, Yejin Bang, Mustafa Shukor, Willy Chung, Jade Yu, Allen Bolourchi, Theo Moutakanni, and Pascale Fung. Action100m: A large-scale video action dataset. *arXiv preprint arXiv:2601.10592*, 2026.

[11] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

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

[13] 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.

[14] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[15] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

[16] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[17] 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.

[18] Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world. *arXiv preprint arXiv:2601.15282*, 2026.

[19] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[20] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[21] Google. Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. Google Developers Blog, 2025. URL `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`. Accessed: 2025-05-06.

[22] Google. Nano banana 2: Google's latest AI image generation model. Google Blog, feb 2026. URL `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`. Accessed: 2026-05-06.

[23] Google DeepMind. A new era of intelligence with Gemini 3, November 2025. URL `https://blog.google/products-and-platforms/products/gemini/gemini-3/`. Accessed: 2026-05-05.

[24] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025.

[25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[26] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

[27] Kunyang Han, Yibo Hu, Mengxue Qu, Hailin Shi, Yao Zhao, and Yunchao Wei. Rose: Revolutionizing open-set dense segmentation with patch-wise perceptual large multimodal model, 2024. URL `https://arxiv.org/abs/2412.00153`.

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

[29] Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.

[30] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[31] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025.

[32] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[33] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[34] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[36] Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025.

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

[38] Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025.

[39] Baijiong Lin, Weisen Jiang, Yuancheng Xu, Hao Chen, and Ying-Cong Chen. Parm: Multi-objective test-time alignment via preference-aware autoregressive reward model. *arXiv preprint arXiv:2505.06274*, 2025.

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[41] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

[42] 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*, pages 216–233. Springer, 2024.

[43] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

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

[45] Ge Ya Luo, Gian Mario Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond fvd: Enhanced evaluation metrics for video generation quality. *arXiv preprint arXiv:2410.05203*, 2024.

[46] Yinyi Luo, Wenwen Wang, Hayes Bai, Hongyu Zhu, Hao Chen, Pan He, Marios Savvides, Sharon Li, and Jindong Wang. Torchumm: A unified multimodal model codebase for evaluation, analysis, and post-training. *arXiv preprint arXiv:2604.10784*, 2026.

[47] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

[48] Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.

[49] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

[50] OpenAI. Introducing our latest image generation model in the API. OpenAI Blog, apr 2025. URL `https://openai.com/index/image-generation-api/`. Accessed: 2025-05-06.

[51] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[52] Jie Qin, Jiancheng Huang, Limeng Qiao, and Lin Ma. Star: Stacked autoregressive scheme for unified multimodal learning. *arXiv preprint arXiv:2512.13752*, 2025.

[53] 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.

[54] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[56] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[57] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[58] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29029–29039, 2025.

[59] Zhongwei Ren, Yunchao Wei, Xiao Yu, Guixun Luo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld 2: Learning transferable knowledge from real-world videos. *arXiv preprint arXiv:2602.10102*, 2026.

[60] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

[61] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[62] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

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

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

[65] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[66] Wentao Shi, Xiangnan He, Yang Zhang, Chongming Gao, Xinyue Li, Jizhi Zhang, Qifan Wang, and Fuli Feng. Enhancing long-term recommendation with bi-level learnable large language model planning. *arXiv preprint arXiv:2403.00843*, 2024.

[67] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[68] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

[69] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

[70] Boming Tan, Xiangdong Zhang, Ning Liao, Yuqing Zhang, Shaofeng Zhang, Xue Yang, Qi Fan, and Yanyong Zhang. Dreamworld: Unified world modeling in video generation. *arXiv preprint arXiv:2603.00466*, 2026.

[71] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *arXiv preprint arXiv:2503.20752*, 2025.

[72] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL `https://arxiv.org/abs/2302.13971`.

[74] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[75] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[76] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[77] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. 2017.

[78] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[79] Jiacong Wang, Bohong Wu, Haiyong Jiang, Zhou Xun, Xin Xiao, Haoyuan Guo, and Jun Xiao. World to code: Multi-modal data generation via self-instructed compositional captioning and filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4608–4623, 2024.

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

[81] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[82] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025.

[83] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

[84] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, Yitao Liang, and Team CraftJarvis. Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34153–34189, 2023.

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

[86] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL `https://arxiv.org/abs/2508.02324`.

[87] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[88] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. 37:68082–68119, 2024.

[89] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[90] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

[91] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[92] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

[93] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[94] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[95] Cheng Yang, Haiyuan Wan, Yiran Peng, Xin Cheng, Zhaoyang Yu, Jiayi Zhang, Junchi Yu, Xinlei Yu, Xiawu Zheng, Dongzhan Zhou, et al. Reasoning via video: The first evaluation of video models' reasoning abilities through maze-solving tasks. *arXiv preprint arXiv:2511.15065*, 2025.

[96] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

[97] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2376–2385, 2025.

[98] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[99] 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. *arXiv preprint arXiv:2308.02490*, 2023.

[100] Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Ar-grpo: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924*, 2025.

[101] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

[102] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

[103] Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

Appendix

In this supplementary appendix, we provide additional details and analyses:

- Appendix A covers implementation details of the two-stage training pipeline
- Appendix B presents further analysis of the VR-X benchmark and evaluation metrics
- Appendix C includes additional ablation studies and visualizations.

A Implementation Details

A.1 Cold Initialization

Data Construction. We construct our SFT training data from multiple sources, as summarized in Tab. A.2. These datasets vary in size and diversity; we weight all subsets according to their sample counts during training. Using the pipeline shown in Fig. A.1, we filter raw data into curated training samples. This data processing pipeline consists of four stages. First, we select raw video sequences from the source data and perform scene-aware temporal sampling via PySceneDetect [9] at 0.27 FPS, which preserves richer information than random sampling. This is followed by SigLIP2-based [75] deduplication and VLM-based filtering to remove low-quality frames such as facial close-ups, blurred frames, and blank screens, typically yielding hundreds of frames from a minute-long video. Second, we leverage Qwen3.5-397B [54] to synthesize reasoning-oriented questions and corresponding textual answers from the sampled sequences, with approximately 10 key steps per trajectory. Third, conditioned on these QA pairs, we prompt Qwen3.5 to identify the most relevant query image and key-step frames. Finally, we apply a rigorous quality filter to discard low-quality samples, including those with image-text mismatches, trivial questions requiring no reasoning or planning, and poor visual quality. This pipeline filters out nearly 80% of the candidate pool, ensuring that the remaining samples represent the most informative and logically sound visual reasoning trajectories.

For non-video sources such as VisualCoT and ZebraCoT, which are already formatted as image sequences, we bypass temporal sampling and standardize them to match video-derived data. Overall, we curate 310k cold initialization samples from 1.5M raw candidates.

Optimization. For cold initialization, we initialize from Emu3.5-34B and train UniVR using the configurations in Tab. A.1 (second column) on 32 GPUs with full parameters. Each image is resized to 512 on the short side and tokenized by Emu3.5's VQ tokenizer, yielding 1,000–1,500 tokens per image. We cap the maximum sequence length per sample at 15,000 to reduce training overhead while accommodating reasoning trajectories spanning approximately several minutes.

A.2 Reinforcement Learning

Data Construction. UniVR employs a novel visual reasoning reward that requires sequences with complex reasoning procedures. To this end, we filter hard samples from the original 310k training set using the post-initialization model, yielding approximately 3k samples. This subset comprises roughly 2k long-term planning trajectories, predominantly spanning 6–10 reasoning steps—and 1k general reasoning data to preserve task diversity.

Optimization. For reinforcement learning, we initialize from the SFT checkpoint and train with full parameters on 32 GPUs using the configuration in Tab. A.1 (third column). The Qwen evaluator is deployed on an additional 8 GPUs. To improve training efficiency, we use Qwen3-VL-30B-A3B as the evaluator. Larger variants (e.g., Qwen3-VL-235B or Qwen3.5-397B) yield more accurate rewards but significantly increase training latency. Classifier-Free Guidance (CFG) is disabled during rollout.

A.3 Reward Computation and Evaluation Protocol

For an instruction-image pair (q, I), VR-GRPO uses the following procedure.

1. **Generate rollouts.** Sample $K = 8$ visual trajectories $\{y_1, \ldots, y_K\} \sim \pi_\theta(\cdot \mid q, I)$.
2. **Compute global preferences.** Compare candidate trajectory pairs using the VLM evaluator and aggregate Win/Tie/Loss outcomes with Eq. 2. A win contributes 1, a tie contributes $1/2$, and a loss contributes 0; division by $K - 1$ gives $R_g \in [0, 1]$.

Table A.1: Training hyperparameters for Cold-Start and RL stages.

Hyperparameters	Cold-Start	RL
Learning rate	5×10^{-4}	1×10^{-5}
LR scheduler	Cosine	Cosine
Weight decay	0.1	0.1
Gradient norm clip	5.0	5.0
Warm-up steps	700	0
Training steps	10k	250
Sequence length	15000	15000
Batch Size	128	128
Resolution	[512, 640]	[512, 640]
λ	-	2.0

3. **Align variable-length trajectories.** For trajectory lengths L_k, use $T = \min_k L_k$ aligned temporal segments. With one-based frame indices,

$$\mathcal{S}_k(t) = \left\{ \left\lfloor \frac{(t-1)L_k}{T} \right\rfloor + 1, \ldots, \left\lfloor \frac{tL_k}{T} \right\rfloor \right\}, \qquad t = 1, \ldots, T. \tag{4}$$

Average normalized CLIP frame embeddings within each segment:

$$z_k(t) = \frac{1}{|\mathcal{S}_k(t)|} \sum_{i \in \mathcal{S}_k(t)} \Phi(y_{k,i}). \tag{5}$$

This partitions each trajectory without padding or interpolation. Compute the cross-trajectory uncertainty $\sigma(t)$ in Sec. 3.2 and select $t^* = \arg\max_t \sigma(t)$.

4. **Compute local preferences.** Map the selected aligned segment back to each trajectory and compare local windows of $W = 4$ frames around that stage, restricted to valid frame indices. Aggregate local Win/Tie/Loss outcomes by Eq. 2 to obtain $R_s \in [0, 1]$. Uncertainty is a localization proxy, not an error label; multiple valid strategies can also produce divergence.
5. **Fuse rewards.** Compute $R_{\text{reason}} = R_g - \lambda|R_g - R_s|$ with $\lambda = 2$. The disagreement penalty encourages agreement between completion and process assessments; it does not guarantee physical correctness.
6. **Add the format reward.** Let R_{format} be the indicator that the trajectory satisfies the required resolution and reasoning-step constraints. The standard VR-GRPO reward is $R_{\text{total}} = R_{\text{format}} + R_{\text{reason}}$.

Evaluation criteria. The evaluator considers the instruction, candidate visual trajectories, and available reference information. The following is a condensed description of the prompt criteria, not a verbatim production prompt:

> Assess task completion and reasoning coherence, including progress toward the instructed objective and consistent intermediate states; physical plausibility of state transitions and object interactions; and visual consistency of object identity, spatial relationships, and temporal evolution. For a pair of candidates, indicate whether the first is better, comparable, or worse (Win/Tie/Loss).

The comparison protocol follows [82]; TIE outcomes allow equivalent valid solutions without forcing a ranking. In compatibility ablations, HPSv3 frame scores and CLIP image-instruction similarities are averaged over the trajectory and added only when those reward components are enabled.

B Details on VR-X Benchmark

The VR-X evaluation set comprises 1.8k high-quality reasoning trajectories curated by professional annotators from a held-out subset (detailed distribution in Tab. A.2). VR-X employs two metrics: VLM score and JEPA similarity score, detailed below.

VLM Evaluation Detail. For the VLM score, we design a unified prompt covering visual quality, task completion, logical coherence, physical dynamics, and temporal consistency. Both the GT and

Table A.2: Data composition.

Category	Data Source	Frames	Ratios
Visual Guidance	EgoDex [30]	289,053	23.7%
	Action100M [10]	109,029	5.0%
	Epic-kitchen [16]	53,605	2.5%
	VideoCraftBench [59]	2,272	1.0%
Robot Manipulation	AgiBot [7]	427,267	24.5%
	Droid [34]	13,000	1.0%
	Bridge [78]	14,850	1.6%
	ZebraCoT-Robot [36]	54,270	3.5%
Editing	ZebraCoT-Multiobject [36]	128,075	6.5%
Spatial Perception	ThinkMorph-Navigation [24]	68,568	11.2%
	ZebraCoT-Embodiment [36]	58,931	3.2%
Visual Search	VisualCoT [63]	30,000	4.9%
	ThinkMorph-Search [24]	13,980	2.3%
Puzzle & Game	VRBench [95]	13,000	1.0%
	Zebra-Jigsaw [36]	43,798	7.1%
	ThinkMorph-VisPuzzle [24]	13,000	1.0%

generated sequence are fed into the VLM, enabling it to reference the GT logic process and action dynamics for more accurate judgment. To assess the alignment between VLM and human evaluation, we first have professional annotators score model outputs on the benchmark. The samples are then shuffled and presented with corresponding VLM and human scores, allowing annotators to judge which is superior. From this, we compute the Spearman correlation [69] between human and VLM scores, which reaches approximately 0.85, indicating a high degree of human alignment.

B.1 Cross-Family Evaluator Validation

To test sensitivity to the Qwen-family evaluator, we also score generated trajectories with GLM-5 and GPT-5.5 and compare evaluator scores with blind human judgments. Tab. B.3 preserves the ordering of the shown methods across all three families. This reduces, but does not eliminate, concern about evaluator-family preference; correlation with human judgments is not proof of accurate physical reasoning on every example. ZS denotes zero-shot evaluation.

Table B.3: **Cross-family evaluation on VR-X.** Scores are shown for zero-shot baselines and UniVR. The final row reports human correlations from the full evaluator-validation study, not correlations recomputed on these displayed rows.

Method	Qwen3.5	GLM-5	GPT-5.5
Show-o2 (ZS)	22.2	21.5	23.0
Bagel (ZS)	30.8	31.2	30.1
Emu3.5 (ZS)	39.8	40.9	40.5
UniVR	**58.2**	**59.0**	**57.5**
Spearman (human)↑	0.85	0.84	0.89

JEPA Evaluation Detail. For the JEPA score, we follow the implementation in [45]. Specifically, its computation resembles traditional video metrics such as FVD [76], but replaces the I3D [8] feature extractor with a V-JEPA [1] encoder. Image sequences are compressed into 1280-dimensional latent vectors, and the distance between two feature distributions is computed via Maximum Mean Discrepancy (MMD) with a polynomial kernel—smaller distances indicate higher similarity. Since the V-JEPA encoder is trained to encode spatiotemporal coherence and physical dynamics, the JEPA score serves as a complement to the VLM score, assessing whether generated sequences conform to real-world transitions. This metric is applied only to the long-term planning subset, as it targets visual sequences containing real-world dynamics and is unsuitable for single-step reasoning tasks

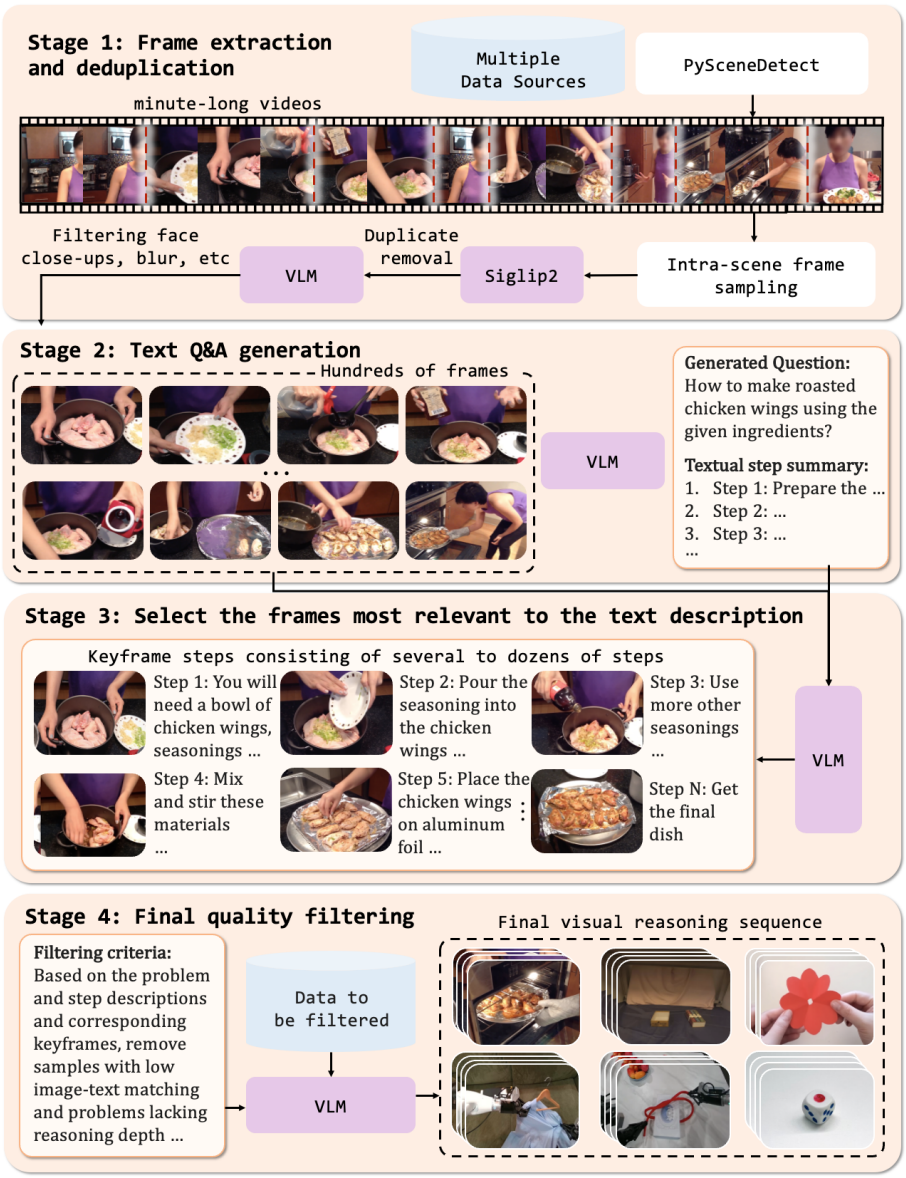

Figure A.1: **Data Generation Pipeline.**

or simulated scenarios in general reasoning. Compared to FVD, the JEPA score converges with approximately 1,000 samples, aligning with the scale of the VR-X benchmark.

C More Analysis and Visualizations

Visual reasoning rewards mitigate reward hacking. Fig. C.2 presents additional cases where the global reasoning reward alone yields near-perfect VLM scores, yet manual inspection reveals physical and logical errors, such as implausible hanger-garment interactions, incorrect wine-pouring dynamics, and flawed towel-replacement logic. These samples span extended durations (30+ seconds), with errors localized to merely a few frames that global VLM assessment easily overlooks. This phenomenon is reflected in the training curves (Fig. C.2, bottom left): the global-reward curve exhibits sharp spikes and top-end oscillations, indicating shortcut-seeking behavior that neglects intermediate-step correctness. In contrast, the VR-GRPO curve (bottom right) ascends more smoothly. Training uses 250 RL steps (Tab. A.1); the plotted curves show only the first 200 steps.

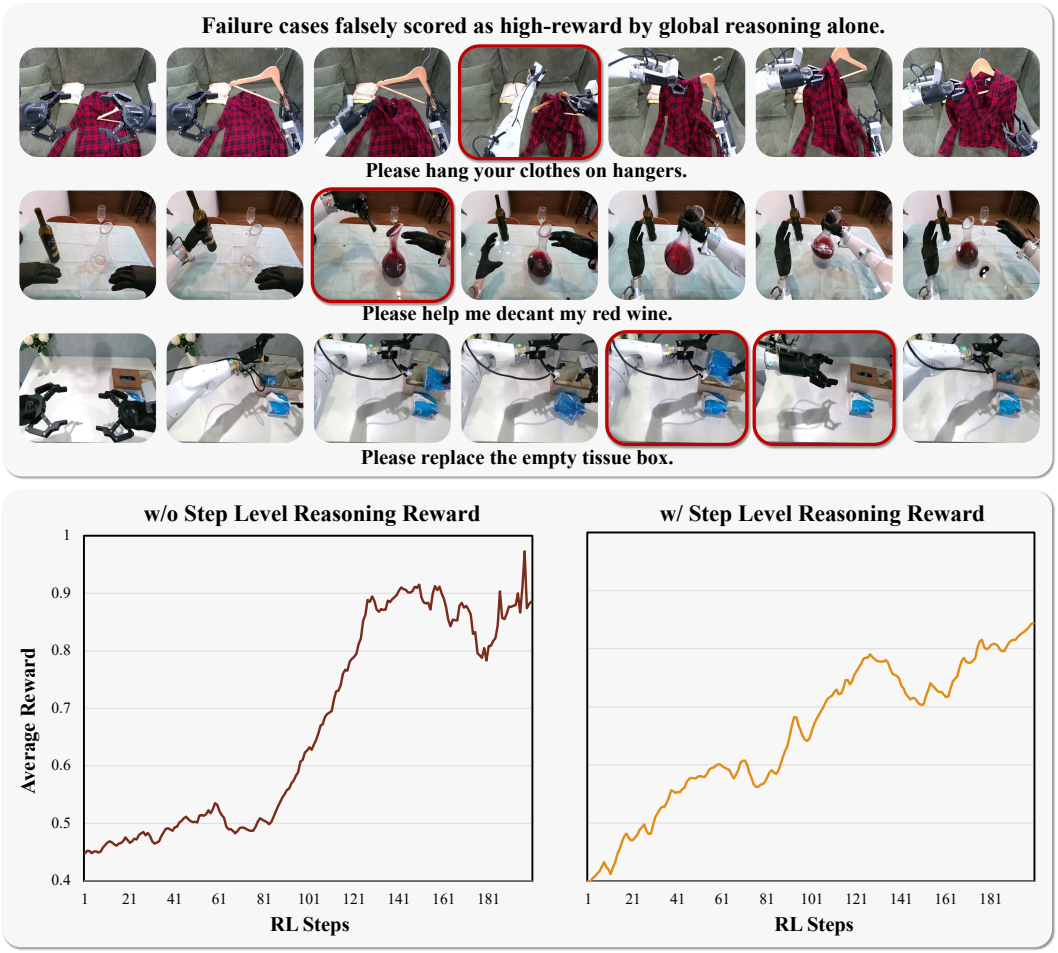

Figure C.2: **Analysis of the Visual Reasoning Reward.** (Up) Relying solely on the global reward misses step-level errors in long-horizon planning tasks, *e.g.* hangers penetrating garments (top row), erroneous liquid pouring dynamics (middle row), and jittering transitions between paper towels (bottom row). (Down) RL reward curves with and without the step-level reward component.

VR-GRPO stabilizes long-horizon visual reasoning. We analyze the scalability of VR-GRPO across varying temporal horizons by partitioning test samples into four duration-based groups: <10s, 10–30s, 30–60s, and >60s. As shown in Tab. C.4, VR-GRPO delivers the most pronounced gains on >30s sequences. We attribute this to its step-level reward design, which explicitly maintains logical coherence and physical consistency at intermediate stages. By correcting error-prone steps, VR-GRPO effectively mitigates compounding errors that typically accumulate during long-range prediction, thereby exhibiting superior stability on extended reasoning traces.

Table C.4: Performance across different time span.

Method	<10s	10-30s	30-60s	>60s
Emu3.5	54.2	48.0	38.9	21.7
Cold-Start	58.7	54.5	41.1	26.9
UniVR	71.1	61.0	56.7	45.6
△*v.s.* Emu3.5	↑ 16.9	↑ 13.0	↑ 17.8	↑ 23.9

More Comparisons with Baselines. Fig. C.4 provides additional side-by-side comparisons with Gemini 3 Pro + Nano Banana 2 and Emu3.5 on identical test samples. While both baselines generate high-fidelity visual appearances, they exhibit execution inaccuracies in fine-grained tasks such as rope manipulation, garment unfolding and folding, paper-folding dynamics, and environmental consistency during policy execution. Benefiting from our visual reasoning training pipeline, UniVR achieves comparable visual fidelity while demonstrating superior logical coherence and physical consistency.

More Visualization. Fig. C.3 shows results of UniVR across additional scenarios, generating long-horizon sequences with coherent logic, physical dynamics, and temporal consistency.



Figure C.3: **More visualization of UniVR.**



Figure C.4: **Comparison with Gemini and Emu3.5.** In each group, the first, second, and third rows correspond to Gemini 3 Pro + Nano Banana 2, Emu3.5, and UniVR, respectively.

C.1 Repeated RL Runs

We repeat RL training three times with different random seeds and the same configuration. Tab. C.5 reports mean and standard deviation across these runs; uncertainty reflects training-seed variability, not a confidence interval or repeated sampling of test examples. The original main-table run is retained separately. Only the configurations explicitly listed here have multi-seed evidence; it should not be attributed to the distinct mixed HPS/CLIP configuration in Tab. 2c.

Table C.5: **Main-result variability.** Mean $\pm$ standard deviation over three independent RL runs.

UniVR evaluation	LP↑	GR↑	Overall↑	JEPA↓
Original run	63.8	55.4	58.2	13.01
Three-run mean	63.1 ± 1.4	56.5 ± 1.3	58.6 ± 1.3	13.03 ± 0.11

Table C.6: **Reward-component variability.** Mean $\pm$ standard deviation over three RL runs for each listed configuration. Pairwise denotes the preference formulation.

R_g	R_s	Pairwise	LP↑	GR↑	JEPA↓
✓			45.5 ± 2.0	45.2 ± 1.8	22.39 ± 0.18
	✓		53.9 ± 1.5	46.5 ± 1.4	15.91 ± 0.14
✓	✓		60.9 ± 1.2	52.4 ± 1.3	13.02 ± 0.10
✓	✓	✓	63.1 ± 1.4	56.5 ± 1.3	13.03 ± 0.11

C.2 Human Validation of Step-Focal

Error localization. We randomly sample 200 training instances from the early stage (first 50 RL steps) and 200 from the late stage (last 50 steps), totaling 3,200 rollout trajectories. Comparing uncertainty-peak regions with human-annotated error locations yields 77.3% error coverage early in training and approximately 65% late in training, when errors are fewer and subtler. These coverage measurements support localization as a useful proxy, but do not establish its precision or exclude divergence from multiple valid solutions.

Reward agreement. A separate analysis uses 200 training examples and 1,600 rollout trajectories independently assessed by PhD-level annotators. Spearman correlation with human judgments is $\rho = 0.76$ for global trajectory rankings and $\rho = 0.89$ for local pairwise preferences. On an additional 100 challenging examples involving knot tying, paper folding, block stacking, and cloth folding, the reported correlations are 0.71 and 0.82, respectively. The global and local analyses assess different scopes, so these correlations should not be interpreted as a matched significance test.

Annotation documentation—author confirmation required. Participant recruitment, the complete annotation instructions, compensation arrangements, and institutional ethics approval or exemption status have not been established in this draft. The authors must verify and document these details before submission; no approval or exemption is asserted here.

C.3 Limitations and Broader Impacts

Training a 34B model on long visual sequences requires substantial compute. UniVR also depends on textual instructions, LLM-assisted data construction, and VLM-based feedback; it is not a language-free learning system. VR-X comparisons remain centered on its source distributions, although adapted baselines and held-out benchmarks provide complementary tests. Cross-family evaluation mitigates, but cannot rule out, evaluator bias, and general-purpose VLMs have limited fine-grained physical grounding. High trajectory variance can also reflect multiple valid strategies rather than an error. Optimal interaction among visual, textual, and auditory reasoning remains open.

Potential benefits include more interpretable visual planning traces and improved multimodal assistance. Risks include misleading or physically implausible generated sequences, unsafe use of such sequences as executable plans, and privacy or licensing concerns when reusing source videos. Generated plans should not be treated as safety-validated robotic actions. Source-specific licenses and

consent restrictions must be respected. Release-specific safeguards and the complete asset-license inventory require author verification; this draft does not claim that these measures have already been implemented.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction distinguish visual intermediate reasoning traces from textual instructions and VLM-based supervision, and delimit the comparison claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

Justification: This is an empirical work and does not present formal theoretical results requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No].

   Justification: Appendix A.3 specifies reward aggregation and temporal alignment, and Appendix A reports training settings. Verbatim production prompts and exact additional evaluator versions remain to be confirmed in this draft.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No].

Justification: Public release is planned, but this draft does not establish the availability of all code, data, weights, and reproduction instructions. The authors must verify release status before submission.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed training and evaluation configurations are provided in the experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix C.1 reports means and standard deviations across three independent RL runs for the main model and listed reward ablations. These characterize training-seed variation, not confidence intervals; other configurations retain point estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: Appendix A reports training and evaluator GPU counts and training steps, but GPU models, memory, and wall-clock costs are not fully documented and require author confirmation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly adhere to all relevant guidelines and regulations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Appendix C.3 discusses beneficial applications, misleading visual outputs, unsafe execution of generated plans, privacy, and licensing risks.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No].

    Justification: Appendix C.3 discusses misuse risks, but release-specific safeguards have not been established in this draft and must be verified by the authors.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No].

    Justification: Existing data, models, and software are cited, but a complete asset-license inventory and redistribution conditions are not yet documented; citations alone do not establish license compliance.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No].

    Justification: The paper describes the model and dataset construction, but accompanying public-release documentation has not been verified in this draft.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No].

    Justification: Human annotations are used in Appendices B.1 and C.2, but complete participant instructions and compensation details require author confirmation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No].

Justification: The institutional review or exemption status and its applicability to the human annotation procedures must be confirmed by the authors. This draft does not assert approval or exemption.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Detailed LLM usage descriptions are included.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.