VIDEO REASONING GENERALIZES TO PERCEPTION, EMBODIMENT, AND REAL-WORLD SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Progress in video generation has been driven largely by visual fidelity. Currently, these models are being deployed as world models and embodied planners, where success depends on reasoning-like competences rather than on pixels. Whether such competences can be acquired by training and whether they generalize beyond the domain on which they were trained remains unanswered. We introduce **XVReason** and **XVReason-Bench**, a video corpus that pairs training data with matched evaluation across four domains: abstract visual reasoning, low-level perception, embodied manipulation, and world simulation. On this foundation, we run a 4×4 cross-domain study on two video backbones. Generalization turns out to be *asymmetric*. Abstract reasoning is the only fine-tuning domain that improves all four task domains on both backbones, with gains of +42.3/+20.6/+10.7% on the three application domains of Wan2.2. This generalization ability also works in different visual domains. Follow-up experiments show that generalization varies with the adaptation regime, the data scale, the denoising stage, and the task composition, and trace the gain to three competences: instruction following, goal completion, and scene stability. An expert-loading ablation recovers most of the reasoning gain through the high-noise expert. These findings point to a recipe, **ReasonGraft**, which blends a reasoning LoRA with an in-domain LoRA at a 3:7 ratio in weight space. It beats in-domain fine-tuning on all three application domains. Video Reasoning is an under-used general-purpose resource for training video generation models.

1 INTRODUCTION

Video generation models have evolved from tools for synthesizing visually appealing sequences into potential learners and reasoners about the physical world (Brooks et al., 2024; Wiedemer et al., 2025). Recent benchmarks reveal latent capabilities in abstract pattern completion, physical prediction, and spatial reasoning, yet these abilities remain fragile across domains (Wang et al., 2026b; Bansal et al., 2025; Liu et al., 2025). This matters because embodied planners, simulators, and interactive world models require rule adherence, dynamic consistency, and instruction following beyond pixel quality. It raises a question the field has not answered: how should a video generation model be trained to acquire abilities to solve these problems?

Two lines of work leave this question open. Inference-time interventions elicit reasoning-like behavior through visual chain-of-thought or thinking-with-video (Huang et al., 2026a; Tong et al., 2026). Training-time efforts use supervised fine-tuning, reinforcement learning with verifiable rewards, or embodied fine-tuning to improve generation correctness (Yang et al., 2025; Chen et al., 2026d; Zhu et al., 2026; Feng et al., 2025), but target specialization within individual domains rather than generalization across them. In language models, code data can enhance the general reasoning capability (Ma et al., 2024), but there is no analog for video generation. Which post-training data yields capability that generalizes has therefore been neither measured nor answered. A fuller discussion of related work is given in Appendix A.

We build the resources that make the question measurable. XVReason is a large video corpus spanning abstract visual reasoning, low-level perception, robotic manipulation, and world simulation, with matched training and evaluation splits for each domain. To our knowledge, XVReason-Bench

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

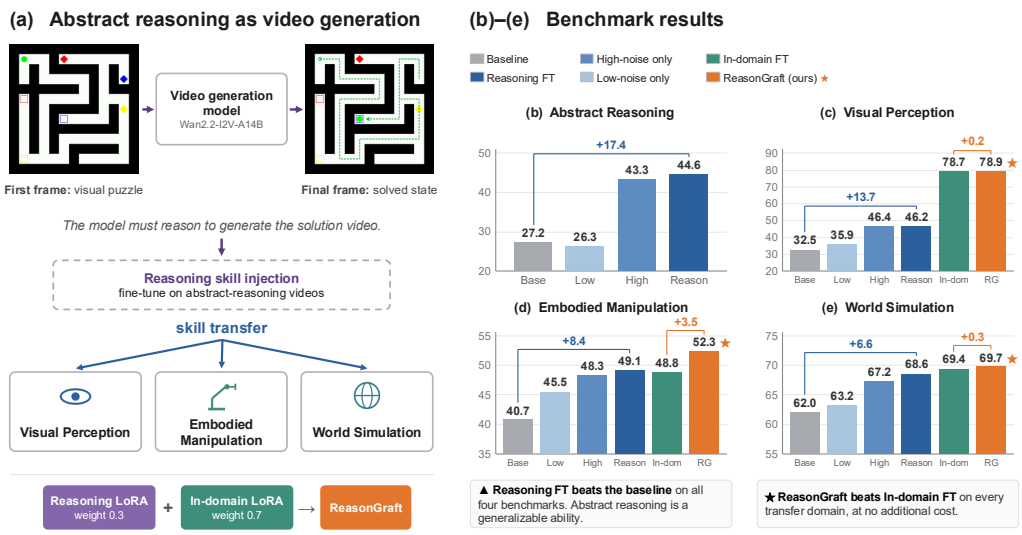

Figure 1: **(a)** An abstract-reasoning video poses visual puzzles. Fine-tuning on such clips generalizes to three unrelated domains. **(b–e)** Reasoning fine-tuning beats the baseline in all four domains. High/Low-noise only: the Reasoning FT LoRA loaded into only the high- or low-noise expert of Wan2.2. ReasonGraft, a 3:7 weight-space blend of the reasoning and in-domain LoRAs.

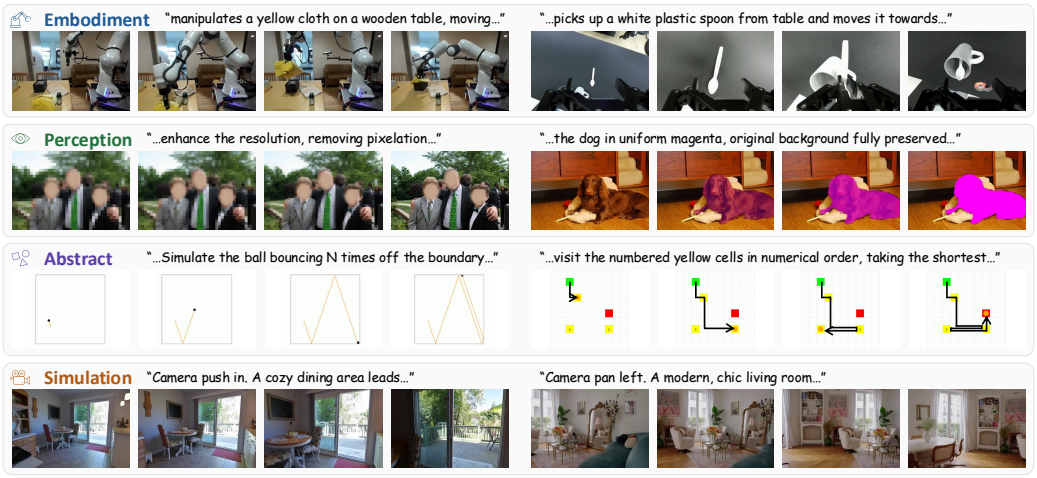

Figure 2: XVReason pairs a training split with a matched benchmark in each of four domains spanning static visual fidelity to multi-step dynamic rollout.

is the first benchmark to evaluate these four competences under a unified protocol. A visual-shift split re-renders the reasoning benchmarks while preserving each task and its solution. With sample budgets, optimization, and inference fixed across domains, we post-train two backbones (Wan2.2-I2V-A14B and LTX-2.3) separately in each domain and evaluate every resulting model in all four domains, yielding a complete 4×4 cross-domain matrix for each backbone (Figure 4). We find that abstract reasoning generalizes along both axes, to the other task domains and to shifted visual domains. The transferable effect is carried mainly by the high-noise expert, and its strength depends on the adaptation regime, the data scale, and the task composition of the training data. Our contributions are:

(i) **XVReason and XVReason-Bench**, a paired training–evaluation suite over abstract reasoning, perception, embodiment, and simulation, with a task-preserving visual-shift split.

(ii) **Asymmetric generalization**: abstract reasoning is the only training source that improves all four task domains on both backbones and yields the strongest mean cross-domain transfer; adaptation regime, scale, denoising stage, and task composition govern its strength, and its robustness is examined under visual domain shifts.

(iii) **ReasonGraft**, a 3:7 weight-space blend of reasoning and in-domain low-rank adapters (LoRAs) that outperforms in-domain fine-tuning on every application-domain benchmark, without additional training cost (Figure 1).

2 XVREASON AND XVREASON-BENCH

A general-purpose world model must close the loop from symbolic cognition to physically grounded control. Following analyses of video world models (Wiedemer et al., 2025; Wang et al., 2026b; Deng et al., 2026; Duan et al., 2025), we identify four major capabilities. **Abstract reasoning** is the discovery and application of logical rules from symbolic visual patterns. **Perception** is the resolution of fine-grained visual structure at the pixel level. **Embodiment** is the generation of physically plausible robot actions that complete a real-world task. **Simulation** is the rollout of a spatio-temporally consistent visual world. The four span a spectrum from static fidelity to multi-step dynamic rollout. To train and evaluate along all four axes, we build *XVReason* and *XVReason-Bench*: each domain contributes a training split and an evaluation suite (Table 1). Dataset samples are shown in Figure 2.

2.1 FOUR TASK DOMAINS

Abstract reasoning. This domain draws on five works: VBVR (Wang et al., 2026b), VIPER (Li et al., 2026), Gen-ViRe (Liu et al., 2025), V-ReasonBench (Luo et al., 2025), and VideoThinkBench (Tong et al., 2026). The training clips are sampled from the VBVR dataset, which covers 100 parameterized task types that span visual rule application, multi-step problem solving, and physical prediction (Appendix B); evaluation uses all five benchmarks. Because VBVR-Bench is paired with the training distribution, it receives half the weight, and the other four benchmarks share the remaining half equally:

$$S_{\text{reason}} = \tfrac{1}{2}\, S_{\text{VBVR}} + \tfrac{1}{8}\big(S_{\text{VIPER}} + S_{\text{Gen-ViRe}} + S_{\text{V-Reason}} + S_{\text{VideoThink}}\big). \tag{1}$$

All five scores lie in $[0, 100]$. Appendix E specifies the evaluation subsets and scoring details.

Visual perception. Following Wiedemer et al. (2025), we cast six low-level vision problems as image-to-video tasks over COCO 2017 (Lin et al., 2014). Three are restoration tasks (*super-resolution*, *denoising*, *low-light enhancement*) and the other three are annotation-extraction tasks (*instance segmentation*, *edge detection*, *keypoint localization*). This yields 177,407 training clips, in each of which the input eases into the target and then holds it statically. The benchmark holds out 300 instances, 50 per task, over COCO images disjoint from all training pools.

Scoring the last frame alone conflates a model that never produces the target with one that produces it and then drifts away. We therefore sample 8 frames uniformly from the second half of the clip, score each against the ground truth, and keep the maximum:

$$s_t(v) \;=\; \max_{k=1,\dots,8} m_t\big(f_{i_k}, G\big), \qquad i_k \;=\; \left\lceil \tfrac{T}{2} \right\rceil + \left\lfloor \tfrac{(k-1)(T-\lceil T/2 \rceil)}{7} \right\rfloor, \tag{2}$$

where T is the clip length, m_t the metric of task t, and G the target image. The three annotation-extraction tasks are scored against their rendered targets by mIoU, a tolerance-matched edge F-measure, and a diagonal-normalized keypoint distance; the three restoration tasks share one score that combines reference-based perceptual fidelity with image quality. The per-task metrics are in Appendix E. The six per-task scores are rescaled to $[0, 100]$ and averaged.

Embodied manipulation. Training data combines a real and a simulated source. DROID (Khazatsky et al., 2024) contributes 94,237 clips of real teleoperated manipulation from 13 academic labs, re-captioned by Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) into third-person descriptive sentences, and RoboTwin 2.0 (Chen et al., 2026c) contributes 94,608 clips from a procedurally scripted dual-arm simulator covering 50 task templates. We mix in a further 100,000 clips filtered from the

Table 1: Details of XVReason and XVReason-Bench. Durations are means over the training split with 95% confidence intervals. The main results reported in each domain were trained on 50K data.

Domain	Clips	Avg. duration (s)	Evaluation set	Score
Abstract	500,000	3.64 ± 0.03	5 benchmarks	weighted benchmark mean
Perception	177,407	3.06	300	6-task mean
Embodied	288,845	4.36 ± 0.04	250	VLM rubric
Simulation	62,446	6.40 ± 0.11	1,000	7 aspects, two protocols
Total	1,028,698			

3.03M-clip RoVid-X corpus released with RBench (Deng et al., 2026); the filtering protocol is in Appendix B. We evaluate on the task-oriented subset of RBench: 250 keyframe–prompt pairs spread evenly over five task families, and report RBench's Task Completion score, rescaled from its five-point scale to $[0, 100]$. Following RBench, Task Completion is the mean of two components judged by GPT-5 (OpenAI, 2025) from a temporal grid of uniformly sampled frames with a VQA-style checklist (Appendix E).

World simulation. We draw 62,446 clips from SpatialVID-HQ (Wang et al., 2026a) and RealCam-Vid (Zheng et al., 2025b), the latter supplying RealEstate10K (Zhou et al., 2018) indoor walkthroughs and DL3DV-10K (Ling et al., 2024) orbits, keeping only static-scene clips. Captions are rewritten into a fixed camera-motion template. We evaluate on the static $\times$ photorealistic subset of WorldScore (Duan et al., 2025): 1,000 instances, half indoor, half outdoor, each with a starting image, a camera path, and a prompt list, scored on its seven aspects.

The simulation axis is reported under two protocols. The **absolute** protocol is the uniform mean of the seven aspects. For comparisons within a backbone, we introduce a **calibrated** protocol that emphasizes the world-simulation abilities the domain requires and keeps a nearly static clip from scoring high: the camera score passes through a logistic centered on the backbone's unadapted camera score T_{base}, and the four *ability* aspects are weighted against the three *consistency* aspects 70:30 (Appendix E). The rest of the paper uses the calibrated protocol, whose scores are comparable only within a backbone.

2.2 Visual domain shifts

XVReason-Bench also tests visual domain generalization, which asks whether its advantage persists when the same reasoning problem is presented in a different visual form. Inspired by image transformations in Geirhos et al. (2026), we edit appearance while preserving the task constraints and solution.

For VIPER, Gen-ViRe, V-ReasonBench, and VideoThinkBench, Nano Banana 2 (Google DeepMind, 2026b) edits the first frame, followed by checks of the answer-bearing content. Evaluation uses tasks whose answers remain assessable through VLM-based scoring after the edit. For VBVR-Bench, we connect the procedural generators to Blender to render corresponding objects and scenes in 3D. Figure 3 illustrates both routes, and Appendix F details the transformations and scoring.

The five shifted benchmark scores are combined with the weights of Equation 1. Shifted V-ReasonBench covers six of its 13 tasks and shifted VideoThinkBench two of its four families, both scored by a VLM judge, so shifted scores compare models within the shifted setting and are not comparable to the original benchmark scores.

3 Experimental Setup

All fine-tuning runs share the backbones, training setup, inference settings, and evaluation protocol below.

Backbones. We use **Wan2.2-I2V-A14B** (Wan Team et al., 2025; Wan Team, 2025) and **LTX-2.3-I2V** (HaCohen et al., 2024; 2026). Unless otherwise stated, all results in the main text are on Wan2.2. LTX-2.3 results are summarized in §4.1 and reported in full in Appendix H.

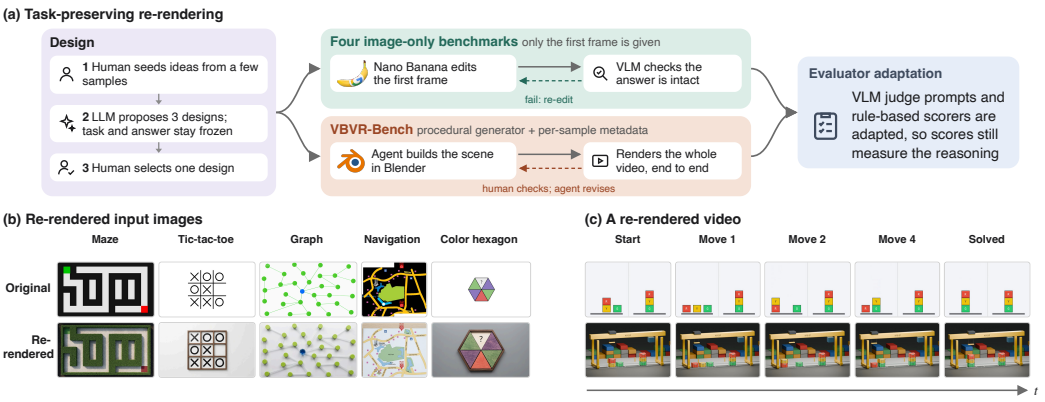

Figure 3: Visual domain shifts preserve the reasoning task while changing its visual presentation. **(a)** The re-rendering pipeline. **(b)** Original and re-rendered input images from the image-only benchmarks. **(c)** Five frames of a VBVR-Bench video and its Blender re-rendering.

Training setup. The default adaptation is LoRA (Hu et al., 2022) with rank 32 and $\alpha = 32$, applied to the attention projections (`q`, `k`, `v`, `o`) on both backbones and, on Wan2.2, also to the feed-forward layers (`ffn.0`, `ffn.2`). Training uses a learning rate of 1×10^{-4} (1×10^{-5} for full fine-tuning) with a global batch size of 32 on 32 NVIDIA A100 GPUs. Unless otherwise stated, all results are obtained under LoRA. Prior work shows that adapting a video model within a single domain is data-efficient, with budgets of tens of thousands of clips or fewer (Zhang et al., 2026; Chen et al., 2026b). Balancing this against the sizes of the four corpora, we sub-sample each corpus to $N = 50{,}000$ clips for the main results, each sampled uniformly across its task types or sub-sources (Appendix B).

Inference. Generation settings are held fixed within a backbone and differ only in the adapter. Both backbones sample by flow matching (Lipman et al., 2023) with a first-order Euler integrator, with classifier-free guidance (CFG) (Ho & Salimans, 2021) against a fixed negative prompt, and every prompt is generated once with a fixed seed. Wan2.2 generates at 16 fps with 40 denoising steps and a static timestep shift of 5.0; LTX-2.3 uses a resolution-adaptive shift. Per-benchmark resolution, clip length, frame rate, and guidance scale are listed in Appendix D.

Evaluation protocol. All four domain scores are rescaled to a common range with higher being better. In what follows, we report Δ relative to the unadapted backbone, $\Delta_{d \to b} = \big(s_b(\theta_d) - s_b(\theta_0)\big)/s_b(\theta_0)$, where θ_d is the model adapted in domain d, θ_0 the baseline, and s_b the score on benchmark b. Relative changes are computed separately for each domain against the same backbone. Evaluation uses a fixed seed and one generation per prompt on the benchmark splits or fixed subsets specified in Appendix E.

4 Findings

4.1 Asymmetric generalization of task solving abilities

Training on abstract reasoning data generalizes across domains. Figure 4 reports the full 4×4 matrix of Δ over baseline; the underlying absolute scores are in Appendix H. The abstract-reasoning row improves perception by +42.3%, embodied by +20.6%, and simulation by +10.7%. Against each domain's own specialist, Reasoning FT matches or exceeds In-domain FT on embodied manipulation. On perception and simulation, it falls short of In-domain FT but stays well above the baseline.

The generalization is asymmetric. Only abstract-reasoning training improves all four task domains on both backbones. It also yields the largest mean off-diagonal Δ: +24.6% on Wan2.2 and +17.0% on LTX-2.3. Every application-domain source, by contrast, reduces at least one other domain. Perception training lowers embodied and simulation scores on both backbones by up to



(a) Wan2.2-I2V-A14B (b) LTX-2.3-I2V

Figure 4: The 4×4 cross-domain matrices (Δ over the base model) on both backbones. The simulation column uses Simulation-Calibrated (§2).

Model	VBVR	VIPER	Gen-ViRe	V-Reason	VideoThink	Average
Wan2.2						
Baseline	30.1	5.0	27.4	1.6	12.9	20.9
Reasoning FT	38.0	7.0	39.9	4.7	10.3	26.7
LTX-2.3						
Baseline	28.7	2.0	27.4	1.0	12.6	19.7
Reasoning FT	33.0	4.0	35.7	4.7	18.1	24.3

Table 2: Visual domain generalization across the five reasoning benchmarks. Models are evaluated on the same shifted inputs within each benchmark.

-37.2% on LTX-2.3; simulation training lowers reasoning on both (-11.2% and -8.5%); and embodied training lowers simulation on LTX-2.3 (-11.1%). Where these sources do transfer, the gains are modest: their mean off-diagonal Δ is at most +9.9% on Wan2.2 and +4.1% on LTX-2.3. The two complete matrices have a Pearson correlation of 0.96, with absolute scores in Table 7.

4.2 Visual domain generalization

We apply the abstract-reasoning adapter without further training to the visual-shift split of §2.2 and compare it with the base model on identical transformed inputs. Table 2 reports the five benchmark scores and their weighted average, following Equation 1. Reasoning FT raises the average from 20.9 to 26.7 on Wan2.2 and from 19.7 to 24.3 on LTX-2.3. The adapter improves all five benchmarks on LTX-2.3 and four of five on Wan2.2. This comparison measures the aggregate advantage of reasoning adaptation on shifted visual inputs.

5 Factors Governing Generalization

The results above show that abstract reasoning data *can* generalize across domains. To further explore the elements that affect this generalization ability, we characterize this dependence along four axes: adaptation regime, training data scale, denoising stage, and task composition, and identify the conditions under which generalization survives and under which it is destroyed.

5.1 Adaptation regime

Does stronger adaptation buy stronger generalization? For this comparison, we report the two halves of VBVR-Bench separately (Wang et al., 2026b): the in-domain (ID) split, whose 50 task types appear in the training corpus, and the out-of-domain (OOD) split of 50 held-out VBVR task types. Full fine-tuning on XVReason exceeds LoRA *in-domain*: on the ID split it gains +72.6% over the base model against LoRA's +64.5% (Figure 5a), which shows that the full fine-tuning run is well

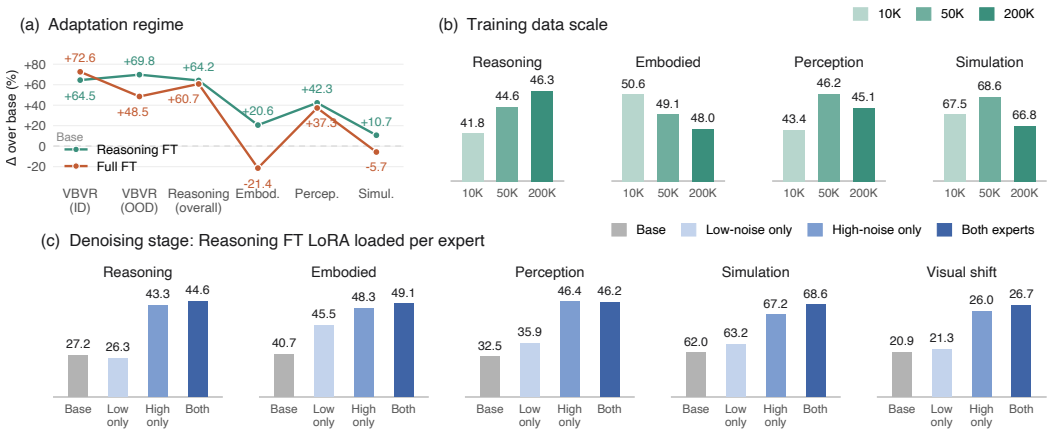

Figure 5: What governs generalization. **(a)** Adaptation regime: relative Δ over the base model after LoRA vs. full FT on XVReason (50K). **(b)** Data scale: scores after LoRA fine-tuning on 10K/50K/200K clips of XVReason. **(c)** Denoising stage: the Reasoning FT LoRA loaded into only the low-noise expert, only the high-noise expert, or both experts of Wan2.2.

optimized rather than under-trained. *Out of domain*, however, its gains shrink sharply on the OOD split and on perception, and on embodied manipulation and simulation the sign flips outright. In-domain performance and cross-domain generalization *dissociate* along the adaptation regime.

A plausible explanation is that full fine-tuning is free to rewrite pretrained representations to fit the target domain, at once enabling stronger specialization and stronger forgetting. The low rank and small magnitude of a LoRA update constrain the shift, so new capability is added on top of what was there rather than over it.

5.2 Data scale

Does more training data buy stronger generalization? We sweep abstract-reasoning training at 10K, 50K, and 200K clips on Wan2.2 (Figure 5b). The reasoning score rises from 41.8 at 10K to 44.6 at 50K and 46.3 at 200K. Mean cross-domain transfer, however, peaks at 50K, even as reasoning performance continues to improve. This motivates an intermediate adaptation strength for cross-domain benefit and anticipates why a partial blend can outperform either endpoint.

5.3 High-noise and low-noise adaptation

Wan2.2 uses separate high-noise and low-noise denoisers. We isolate their contribution by loading the same Reasoning FT checkpoint into the high-noise expert only, the low-noise expert only, or both experts, and compare with the base model. The expert without a LoRA retains its pretrained weights and still executes its part of the noise schedule; prompts, seeds, and sampling settings are unchanged.

Figure 5c compares the four domain scores and the visual-shift score. High-noise-only adaptation recovers 92.4% of the full adapter's gain in the reasoning score and retains most of its application-domain benefit. Low-noise-only adaptation leaves the reasoning score at the base level (26.3 vs. 27.2). The main transferable effect is therefore concentrated in the high-noise expert's update.

5.4 Task composition

Abstract-reasoning tasks ask for two things: inferring a target state, and executing an explicitly stated instruction. We split the 100 training task types by whether the prompt fully specifies the answer, giving 47 reasoning and 53 instruction-following types, and fine-tune Wan2.2 on an equal number of clips from each (Appendix G). The instruction-following subset leads on all three application domains (perception 47.1 vs. 41.7, embodied 49.9 vs. 47.4, simulation 68.5 vs. 64.5). The

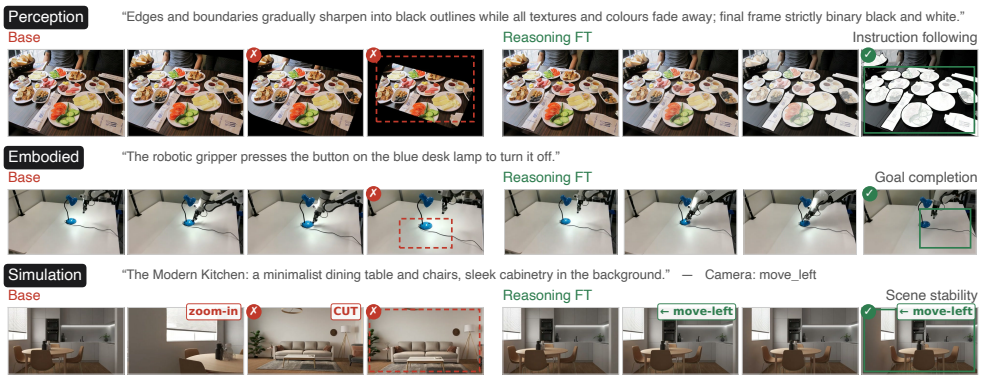

Figure 6: Baseline (red) vs. Reasoning FT (green) on Wan2.2, one case per competence.

two components train largely independent capabilities. Instruction following generalizes more, but reasoning alone still beats the baseline on all three, so reasoning itself generalizes.

6 Why video reasoning can generalize

Across all four domains, the reasoning-tuned model differs from the base model along three competences, classified by *what the output must satisfy* rather than by appearance. Figure 6 shows one matched Baseline/Reasoning FT pair per competence. Nine further cases are in Appendix H. The ability video models acquired through reasoning fine-tuning can be classified into three aspects. **Instruction following**: executing the operation the prompt states; the base model returns a photograph instead of the requested edge map, or skips the specified camera move. **Goal completion**: carrying the action through to the specified target state rather than halting at a half-restored image or a lamp reached for but never switched off. **Scene stability**: keeping content off the task target fixed, so no object appears or disappears and no room changes on revisiting a viewpoint. Every reasoning clip demands all three: a puzzle counts as solved only if the stated rule is applied, the final state is reached, and everything else stays put.

The quantitative results agree. The two largest gains fall in the domains scored by goal completion: perception (+42.3%) and embodied manipulation (+20.6%) both ask whether a specified end state is reached. Simulation (+10.7%) is scored on executing a camera path through a static scene, so its gain rests on instruction following and scene stability. None of this costs general generation quality: on VBench (Huang et al., 2024; 2026b), the mean of nine I2V quality dimensions moves by at most 0.2 points on either backbone, a lower dynamic degree offset by small gains elsewhere (Table 22).

7 ReasonGraft

7.1 Motivation

The findings of the previous section specify the construction. The regime finding requires the adaptation to stay parameter-efficient, so the two capabilities must be combined in LoRA weight space (Wortsman et al., 2022; Ilharco et al., 2023) rather than by one full-parameter run on both corpora; the scale finding says the useful adaptation strength is intermediate, so the combination needs a continuously tunable strength; and the task-composition analysis of §5.4 shows that the reasoning and instruction-following components of the corpus train largely independent capabilities, so the grafted LoRA is trained on the full corpus to carry both. A weighted blend of two LoRAs is the simplest form that satisfies all three.

7.2 Method

Let ΔW_{reason} and $\Delta W_{\text{in-dom}}$ be the weight updates induced by the two LoRAs. ReasonGraft forms

$$\Delta W \;=\; \lambda\,\Delta W_{\text{reason}} \;+\; (1-\lambda)\,\Delta W_{\text{in-dom}}, \qquad W \;=\; W_0 + \Delta W. \tag{3}$$

Model	Perception	Embodied	Simulation
Reasoning FT	46.2	49.1	68.6
In-domain FT	78.7	48.8	69.4
ReasonGraft (7:3)	72.9	48.7	69.7
ReasonGraft (5:5)	77.5	52.1	70.0
ReasonGraft (3:7)	78.9	52.3	69.7
LoRA-LEGO	78.4	52.7	70.1
RobustMerge	78.5	51.3	70.4
Mixed-data FT (3:7)	78.3	40.9	69.5

Table 3: ReasonGraft against both endpoints and against dedicated LoRA-merging methods applied to the same two adapters, on Wan2.2. Ratios are reasoning : in-domain; simulation uses Simulation-Calibrated (§2). Underlined cells exceed In-domain FT.

The weighting is applied to the expanded update $\Delta W = BA$ rather than to the factors A and B separately, so that Equation 3 is exactly a convex combination of the two induced updates. On Wan2.2, the blend is applied independently to the high-noise and low-noise expert LoRAs with the same λ. We call the operation ReasonGraft because it is asymmetric: the in-domain LoRA is the stock that the target task requires, and a single reasoning LoRA is grafted onto it at a small fixed weight. The reasoning LoRA is trained once and reused for every target domain.

7.3 Results

A single blending ratio works for all domains. Table 3 sweeps the blending ratio over $\lambda \in \{0, 0.3, 0.5, 0.7, 1\}$ on Wan2.2, where the endpoints $\lambda = 0$ and $\lambda = 1$ are In-domain FT and Reasoning FT. $\lambda = 0.3$ (3:7) is at or near the maximum in *every* column, best on perception and embodied manipulation and second-best on simulation, so we adopt 3:7 as the default. That one ratio serves every target domain is what makes the recipe usable: there is nothing to tune.

ReasonGraft beats both endpoints. At the default 3:7 ratio, ReasonGraft exceeds both In-domain FT and Reasoning FT in every column of Table 3. The margin over In-domain FT is largest on embodied manipulation (52.3 vs. 48.8), where the reasoning component contributes most, and remains positive even on perception (78.9 vs. 78.7), where In-domain FT is strongest. The λ curve peaking at an interior point is the same phenomenon we observed along the data-scale axis. Both reflect the specialization–generalization tension that runs through the whole analysis. The optimal ratio is thus not an incidental hyperparameter but the expected consequence of the analysis. Dedicated LoRA-merging methods do not add to this: merging the same two adapters with LoRA-LEGO (Zhao et al., 2025) or RobustMerge (Zeng et al., 2025) matches the plain weighted sum to within one point in all three domains, so what matters is which adapters are merged and at what ratio, not the merging operator. Nor is the gain recovered by mixing the data instead: a single LoRA trained on the two corpora mixed 3:7 under the same budget only matches In-domain FT on simulation and falls below it on perception and embodied manipulation, so the benefit comes from blending separately specialized adapters in weight space rather than from joint exposure to both corpora.

8 Limitations and Conclusion

Limitations. Two directions remain open. The relation between reasoning and instruction following deserves a finer treatment, where the reasoning demand of the data can be measured and used to predict generalization. The abstract-reasoning training corpus is predominantly synthetic; extending it to reasoning tasks in recorded real-world footage remains a natural next step beyond changes in visual presentation.

Conclusion. We have built XVReason and XVReason-Bench, fine-tuned two video generation backbones in four domains under one controlled protocol, and evaluated every resulting model in every domain. Generalization in video post-training is directional: abstract reasoning is the only fine-tuning domain that improves all four task domains on both backbones, so reasoning data should be treated as a general-purpose training corpus rather than as a specialist dataset for a reasoning benchmark. It teaches instruction following, goal completion, and scene stability. ReasonGraft turns this into a recipe, grafting a single reasoning LoRA onto each in-domain LoRA at one fixed ratio and improving over both endpoints in every domain without additional training.

AI USE STATEMENT

Generative AI tools were used in three ways. First, vision-language and language models generated parts of the training captions: Qwen3-VL-30B-A3B-Instruct wrote the third-person descriptions of the DROID clips, and the simulation captions were rewritten into a fixed camera-motion template (§2). Second, LLM-based assistants helped locate and organize related work; every cited paper was read and checked by the authors. Third, LLM-based assistants helped revise and polish the writing for readability and consistency. Image-generation models produced the first-frame visual transformations, and VLMs verified these edits and scored the applicable reasoning benchmarks. The authors verified all content and take full responsibility for it.

REPRODUCIBILITY STATEMENT

All experiments build on publicly released backbones (Wan2.2-I2V-A14B and LTX-2.3-I2V) and the cited data sources (§2; curation rules in Appendix B). Every fine-tuning run follows the controlled protocol of §3, and ReasonGraft is a post-hoc convex combination of the expanded updates (Equation 3) that requires no training to reproduce. Inference is deterministic, with one fixed-seed generation per prompt and full sampler settings in Table 6, and every scoring protocol is fully specified in §2 and Appendix E. Code, the curated XVReason corpus, XVReason-Bench together with its visual-shift split (the edited first frames and the Blender-rendered clips), and all LoRA checkpoints will be released soon, subject to the licenses of the underlying data sources.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 58757–58791, 2024.

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

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=9D2QvO1uWj.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

Zefan Cai, Haoyi Qiu, Tianyi Ma, Haozhe Zhao, Gengze Zhou, Tingting Liao, Xinyan Velocity Yu, Kung-Hsiang Huang, Ke Wan, Shawn Lin, Parisa Kordjamshidi, Minjia Zhang, Wen Xiao, Jiuxiang Gu, Nanyun Peng, and Junjie Hu. MMGR: Multi-modal generative reasoning benchmark and evaluation. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026. URL https://arxiv.org/abs/2512.14691. Accepted for publication.

Harold Haodong Chen, Disen Lan, Wen-Jie Shu, Qingyang Liu, Zihan Wang, Sirui Chen, Wenkai Cheng, Kanghao Chen, Hongfei Zhang, Zixin Zhang, Rongjin Guo, Yu Cheng, and Ying-Cong Chen. TiViBench: Benchmarking think-in-video reasoning for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11403–11413, 2026a.

Houyuan Chen, Hong Li, Xianghao Kong, Tianrui Zhu, Shaocong Xu, Weiqing Xiao, Yuwei Guo, Chongjie Ye, Lvmin Zhang, Hao Zhao, and Anyi Rao. UniVidX: A unified multimodal framework for versatile video generation via diffusion priors. *ACM Transactions on Graphics*, 45(4):51, July 2026b. doi: 10.1145/3811304. URL https://doi.org/10.1145/3811304.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *International Conference on Machine Learning (ICML)*, 2026c. URL https://icml.cc/virtual/2026/poster/62192.

Xinyan Chen, Ziyu Guo, Renrui Zhang, Dongzhi Jiang, and Hongsheng Li. OpenCoF: Learning to reason through video generation. *arXiv preprint arXiv:2607.08763*, 2026d.

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. In *International Conference on Machine Learning (ICML)*, 2026.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? A comprehensive survey of world models. *ACM Computing Surveys*, 58(3):57:1–57:38, 2025. doi: 10.1145/3746449.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27713–27724, 2025.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Robert Geirhos, Yuxuan Li, Thaddäus Wiedemer, Neha Kalibhat, Zi Wang, Mani Malek, Øyvind Tafjord, Kevin Swersky, Been Kim, and Priyank Jaini. Visual prompt engineering for video models. *arXiv preprint arXiv:2607.25537*, 2026.

Google DeepMind. Gemini 3.8 Flash model card. https://deepmind.google/models/model-cards/gemini-3-8-flash, 2026a.

Google DeepMind. Gemini 3.1 Flash Image (Nano Banana 2). https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-image, 2026b.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Xuming He, Zehao Fan, Hengjia Li, Fan Zhuo, Hankun Xu, Senlin Cheng, Di Weng, Haifeng Liu, Can Ye, and Boxi Wu. RULER-Bench: Probing rule-based reasoning abilities of next-level video generation models for vision foundation intelligence. *arXiv preprint arXiv:2512.02622*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. arXiv:2207.12598.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024.

Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec, and Ziwei Liu. VChain: Chain-of-Visual-Thought for reasoning in video generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 226–250, San Diego, California, United States, 2026a. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.12. URL https://aclanthology.org/2026.findings-acl.12/.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, March 2026b. doi: 10.1109/TPAMI.2025.3633890.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28991–29017. PMLR, 2025. URL https://proceedings.mlr.press/v267/kang25g.html.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, 2021.

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. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Kuaishou Technology. Kling AI. https://klingai.com, 2025. Accessed 2026.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf. Version 0.9.2, 2022-06-27.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Datasets and Benchmarks Track.

Yifan Li, Yukai Gu, Yingqian Min, Zikang Liu, Yifan Du, Kun Zhou, Min Yang, Wayne Xin Zhao, and Minghui Qiu. Beyond the last frame: Process-aware evaluation for generative video reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20393–20409, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.934. URL https://aclanthology.org/2026.acl-long.934/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 2024.

Xinxin Liu, Zhaopan Xu, Ming Li, Kai Wang, Yong Jae Lee, and Yuzhang Shang. Can world simulators reason? Gen-ViRe: A generative visual reasoning benchmark. *arXiv preprint arXiv:2511.13853*, 2025.

Yang Luo, Xuanlei Zhao, Baijiong Lin, Lingting Zhu, Liyao Tang, Yuqi Liu, Ying-Cong Chen, Shengju Qian, Xin Wang, and Yang You. V-ReasonBench: Toward unified reasoning benchmark suite for video generation models. *arXiv preprint arXiv:2511.16668*, 2025.

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help LLMs reasoning? In *International Conference on Learning Representations (ICLR)*, 2024.

OpenAI. OpenAI GPT-5 system card. https://openai.com/index/gpt-5-system-card/, 2025. arXiv:2601.03267.

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

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

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. *AI Open*, 7:208–226, 2026. doi: 10.1016/j.aiopen.2026.07.002.

Xavier Soria, Edgar Riba, and Angel Sappa. Dense extreme inception network: Towards a robust CNN model for edge detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1912–1921, Los Alamitos, CA, USA, 2020. IEEE Computer Society. doi: 10.1109/WACV45572.2020.9093290. URL https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093290.

Xavier Soria, Angel Sappa, Patricio Humanante, and Arash Akbarinia. Dense extreme inception network for edge detection. *Pattern Recognition*, 139:109461, 2023. ISSN 0031-3203. doi: 10.1016/j.patcog.2023.109461. URL https://www.sciencedirect.com/science/article/pii/S0031320323001619.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, Xinchi Chen, Jun Zhao, Xuanjing Huang, and Xipeng Qiu. Thinking with video: Video generation as a promising multimodal reasoning paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41121–41129, 2026.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. https://github.com/Wan-Video/Wan2.2, 2025.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42592–42603, 2026a.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, et al. A very big video reasoning suite. In *International Conference on Machine Learning (ICML)*, 2026b.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

Jialong Wu, Xiaoying Zhang, Hongyi Yuan, Xiangcheng Zhang, Tianhao Huang, Changjing He, Chaoyi Deng, Renrui Zhang, Youbin Wu, and Mingsheng Long. Visual generation unlocks human-like reasoning through multimodal world models. *arXiv preprint arXiv:2601.19834*, 2026.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Yongtao Ge, and Kaipeng Zhang. WorldMark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026. URL https://arxiv.org/abs/2604.21686.

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.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuxue Yang, Shuyao Shang, Jiahe Wang, Zitong Zhou, Liang Tan, Junhan Zeng, Ruizhi Li, Junyan Li, Yu Liu, Xiao Yang, Yong Li, Jun Zhu, Hongsheng Li, Tieniu Tan, Lue Fan, and Zhaoxiang Zhang. WorldExam: Benchmarking world models from apparent appearance to inherent reactivity. *arXiv preprint arXiv:2608.02603*, 2026. URL https://arxiv.org/abs/2608.02603.

Fanhu Zeng, Haiyang Guo, Fei Zhu, Li Shen, and Hao Tang. RobustMerge: Parameter-efficient model merging for MLLMs with direction robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations (ICLR)*, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Ziyu Zhao, Tao Shen, Didi Zhu, Zexi Li, Jing Su, Xuwu Wang, and Fei Wu. Merging LoRAs like playing LEGO: Pushing the modularity of LoRA to extremes through rank-wise clustering. In *International Conference on Learning Representations (ICLR)*, 2025.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025a.

Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. RealCam-Vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025b.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018.

Tinghui Zhu, Sheng Zhang, James Y. Huang, Selena Song, Xiaofei Wen, Yuankai Li, Hoifung Poon, and Muhao Chen. Video models can reason with verifiable rewards. *arXiv preprint arXiv:2605.15458*, 2026.

APPENDIX CONTENTS

A Related work **17**

A.1 Video generation models 17

A.2 Reasoning in video generative models 17

A.3 Reasoning data and visual presentation 17

A.4 World models and embodied video generation 17

B Data curation details **18**

B.1 Abstract reasoning 18

B.2 Perception 18

B.3 Embodied manipulation 18

B.4 Simulation 19

C Dataset and benchmark samples **19**

D Inference settings **19**

E Benchmark scoring details **22**

E.1 Abstract reasoning 22

E.2 Perception per-task metrics 22

E.3 Embodied scoring 23

E.4 Simulation scoring 23

E.5 Simulation aggregation 23

F Visual domain transformation and evaluation **23**

G Reasoning and instruction-following subsets **25**

G.1 Parameter-level analysis of the LoRA updates 25

H Full result tables **27**

A Related work

A.1 Video generation models

Diffusion models (Ho et al., 2020; Rombach et al., 2022) and flow matching with diffusion transformers (Lipman et al., 2023; Peebles & Xie, 2023) are the dominant paradigms for video generation. Both closed products (Brooks et al., 2024; Kuaishou Technology, 2025) and open efforts (Wan Team et al., 2025; Kong et al., 2024; Zheng et al., 2024; HaCohen et al., 2024) have reached strong temporal coherence and visual realism. Adapting these models after pre-training is now routine, through low-rank adaptation (Hu et al., 2022), full fine-tuning, preference tuning, or controllability tuning. Evaluation has advanced in parallel, from temporal consistency and motion quality (Huang et al., 2024) to physical commonsense and intrinsic faithfulness (Bansal et al., 2026; Zheng et al., 2025a). Together, these works show how video generators can be adapted and evaluated within a target domain, but not whether capabilities acquired through post-training generalize across domains.

A.2 Reasoning in video generative models

Wiedemer et al. (2025) showed that a strong video model can perform perception, physical modeling, and manipulation tasks zero-shot. Subsequent work has probed (He et al., 2025; Li et al., 2026; Chen et al., 2026a; Luo et al., 2025) and improved (Huang et al., 2026a; Tong et al., 2026; Wu et al., 2026) these capabilities, largely through inference-time intervention, and a parallel line builds benchmarks for abstract and spatial reasoning (Wang et al., 2026b; Liu et al., 2025; Yang et al., 2025). Within this line, generalization is measured across in-domain and out-of-domain reasoning tasks. We report on this axis as well, and our object of study is generalization *out of* reasoning training and into downstream application domains, namely perception, embodied manipulation, and simulation. The two questions are orthogonal and complementary. In large language models, training on code data is known to improve general reasoning capability (Ma et al., 2024). We provide the first systematic evidence of an analogous effect in video generation models.

A.3 Reasoning data and visual presentation

Video reasoning corpora supply parameterized procedural tasks (Wang et al., 2026b), verifiable puzzles (VideoRLVR), physical trajectories (PhyWorld), video reasoning demonstrations (OpenCoF), and spatial planning tasks (VR-Bench) (Zhu et al., 2026; Kang et al., 2025; Chen et al., 2026d; Yang et al., 2025). Visual prompt engineering (Geirhos et al., 2026) studies how input-image transformations elicit video-model capabilities at inference time. We use task-preserving changes in visual presentation to evaluate whether the advantage learned through abstract-reasoning fine-tuning generalizes across visual domains.

A.4 World models and embodied video generation

Video generators are increasingly treated as world models (Ha & Schmidhuber, 2018; LeCun, 2022; Brooks et al., 2024; Ding et al., 2025) and re-purposed as interactive environments (Bruce et al., 2024; Alonso et al., 2024), as real-world simulators (Yang et al., 2024; Agarwal et al., 2025), and as embodied controllers (Feng et al., 2025). Complementary evaluation efforts have shown that high visual fidelity alone does not guarantee grounded task performance: RBench (Deng et al., 2026), WorldScore (Duan et al., 2025), WorldModelBench (Li et al., 2025), and WorldArena (Shang et al., 2026). WorldScore is the most relevant here. It decomposes world generation into separately measurable axes of controllability and spatio-temporal consistency, and it forms the basis of our simulation axis (§2). All of these advances remain confined to a single task or a single domain at a time. Table 4 summarizes the capability coverage of these benchmarks against ours. Whether cross-domain training, and in particular training on abstract reasoning, generalizes to physically grounded benchmarks has not been investigated.

Table 4: Capability coverage of video-generation benchmarks. A checkmark marks at least one dedicated task or dimension in that category; Perception covers low-level vision tasks and Simulation covers camera-controlled scene rollout. Ours (XVReason-Bench) is the first to cover all six.

Benchmark	**Reasoning**			**Grounded Tasks**		
	Abstract	Spatial	Physical	Percept.	Manip.	Sim.
VBench-2.0 (Zheng et al., 2025a)	✗	✓	✓	✗	✗	✓
Video-Zero-Shot (Wiedemer et al., 2025)	✓	✓	✓	✓	✓	✗
VideoPhy-2 (Bansal et al., 2026)	✗	✗	✓	✗	✗	✗
Gen-ViRe (Liu et al., 2025)	✓	✓	✓	✗	✓	✗
VBVR-Bench (Wang et al., 2026b)	✓	✓	✓	✗	✗	✗
RULER-Bench (He et al., 2025)	✓	✓	✓	✗	✗	✗
TiViBench (Chen et al., 2026a)	✓	✓	✗	✗	✓	✗
MMGR (Cai et al., 2026)	✓	✓	✓	✗	✗	✗
WorldModelBench (Li et al., 2025)	✗	✗	✓	✗	✓	✗
RBench (Deng et al., 2026)	✗	✓	✓	✗	✓	✗
WorldScore (Duan et al., 2025)	✗	✓	✗	✗	✗	✓
WorldMark (Xu et al., 2026)	✗	✓	✗	✗	✗	✓
WorldArena (Shang et al., 2026)	✗	✓	✓	✗	✓	✗
WorldExam (Yang et al., 2026)	✗	✓	✓	✗	✗	✓
Ours	✓	✓	✓	✓	✓	✓

B Data curation details

B.1 Abstract reasoning

Abstract-reasoning training clips are produced by procedural generators covering 100 parameterized task types. The full pool has 500,000 clips, and the 10K, 50K, and 200K settings are sub-sampled from it. The reasoning/instruction-following ablation partitions the 100 VBVR task types into 47 and 53 types respectively (Appendix G). We denote the domain adapters as Reasoning FT, Perception FT, Embodiment FT, and Simulation FT. They use the default 50K training budget unless another budget is explicitly indicated.

B.2 Perception

The three restoration degradations are a 16×16 block-mean mosaic for super-resolution, additive Gaussian noise with $\sigma \sim \mathcal{U}(15, 50)$ for denoising, and gamma darkening with $\gamma \sim \mathcal{U}(2, 5)$ combined with noise for low-light enhancement. Degradation parameters are seeded per image, so the corpus is reproducible. The three annotation targets are a magenta foreground mask, a DexiNed edge map (Soria et al., 2020; 2023), and a black canvas carrying three colored dots on the nose and both shoulders of a person. Per-task clip counts are seg 40,000, edge 30,000, sr 32,090, low-light 32,103, denoise 32,093, keypoint 11,121. In every clip, the input image eases into the target through a quadratic alpha blend over frames 2–32, and the final 17 frames hold the target statically.

B.3 Embodied manipulation

From DROID, we keep only success episodes of 2.7–8.1 s and retain all three views per episode, two exterior and one wrist-mounted. DROID's native annotations are short imperative commands rather than descriptions of the observed behavior, so we re-caption every episode with Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) into a single third-person descriptive sentence, appending a fixed viewpoint marker to wrist-camera clips so that the generator can tell the views apart. RoboTwin 2.0 covers four robot embodiments and is used with its human-authored paraphrases. RoVid-X clips are 30 fps with a minimum side of 688 px. From the 3.03M released clips we keep those that carry a caption and whose length falls in $[49, 209]$ frames ($\approx$ 1.6–7.0 s), drop any whose `short_caption` has cosine similarity ≥ 0.90 to prompt set of our selected clips from DROID and RoboTwin 2.0, gate the rest on decode success, non-frozen content, and exposure, and rank the survivors by technical visual quality down to the best 100,000.

B.4 SIMULATION

SpatialVID-HQ contributes 33,058 clips of long-range dolly, pan, and truck motion; RealCam-Vid contributes 25,259 RealEstate10K indoor walkthroughs and 4,129 DL3DV-10K orbits. The two sources are complementary, with DL3DV supplying the orbital motion that SpatialVID lacks. Clips are kept only if the scene content is static, if their duration falls in $[3, 13]$ s, and if their text does not narrate scene motion. Every caption is then rewritten into the template `"Camera <move>. <scene description>. Everything in the scene remains completely motionless."`. The resulting camera vocabulary has 14 verbs, and 87.1% of clips fall within the 8 verbs the benchmark evaluates; Table 5 gives the full distribution.

Table 5: Camera-verb distribution of the simulation training split (62,446 clips).

Verb	Clips	Share	Verb	Clips	Share
push in	14,299	22.9%	move left	2,335	3.7%
orbit left	9,051	14.5%	tilt up	1,838	2.9%
orbit right	8,768	14.0%	tilt down	1,420	2.3%
pan right	8,139	13.0%	pull out	1,224	2.0%
pan left	8,089	13.0%	crane up	601	1.0%
fixed	3,532	5.7%	crane down	343	0.5%
move right	2,465	3.9%	roll	342	0.5%

C DATASET AND BENCHMARK SAMPLES

Figures 7 and 8 show representative training clips across the four domains as four evenly spaced frames with the training prompt. Faces are pixelated for display, and the fixed static-scene sentence of the perception prompt template is omitted.

D INFERENCE SETTINGS

Both backbones sample by flow matching with a first-order Euler integrator ($x \leftarrow x + v\,\Delta\sigma$), with classifier-free guidance (CFG) against a fixed negative prompt. Wan2.2 uses a static timestep shift of 5.0 and, being a two-expert MoE, switches experts at a fixed boundary in σ. LTX-2.3 uses a resolution-adaptive shift, which at our latent lengths reaches its cap of $\mu = 2.05$, with the terminal σ clamped to 0.1. Every prompt is generated once with a fixed seed. Table 6 lists the per-benchmark settings.

Table 6: Per-benchmark inference settings.

	VBVR-Bench	WorldScore	RBench	Perception
Wan2.2-I2V-A14B				
Resolution	832×480	832×480	832×480	keep aspect ratio and area, stride 16
Frames	81	81	81	33
fps	16	16	16	16
Steps	40	40	40	40
CFG	3.5	3.5	5.0	5.0
Timestep shift	5.0	5.0	5.0	5.0
Expert boundary	0.9	0.9	0.9	0.875
Seed	42	42	42	42
LTX-2.3-I2V				
Resolution	768×512	768×512	832×480	keep aspect ratio, area 512×768, stride 32
Frames	121	121	121	49
fps	24	24	24	16
Steps	30	30	40	40
CFG	3.0	3.0	3.0	3.0
Seed	42	42	42	42

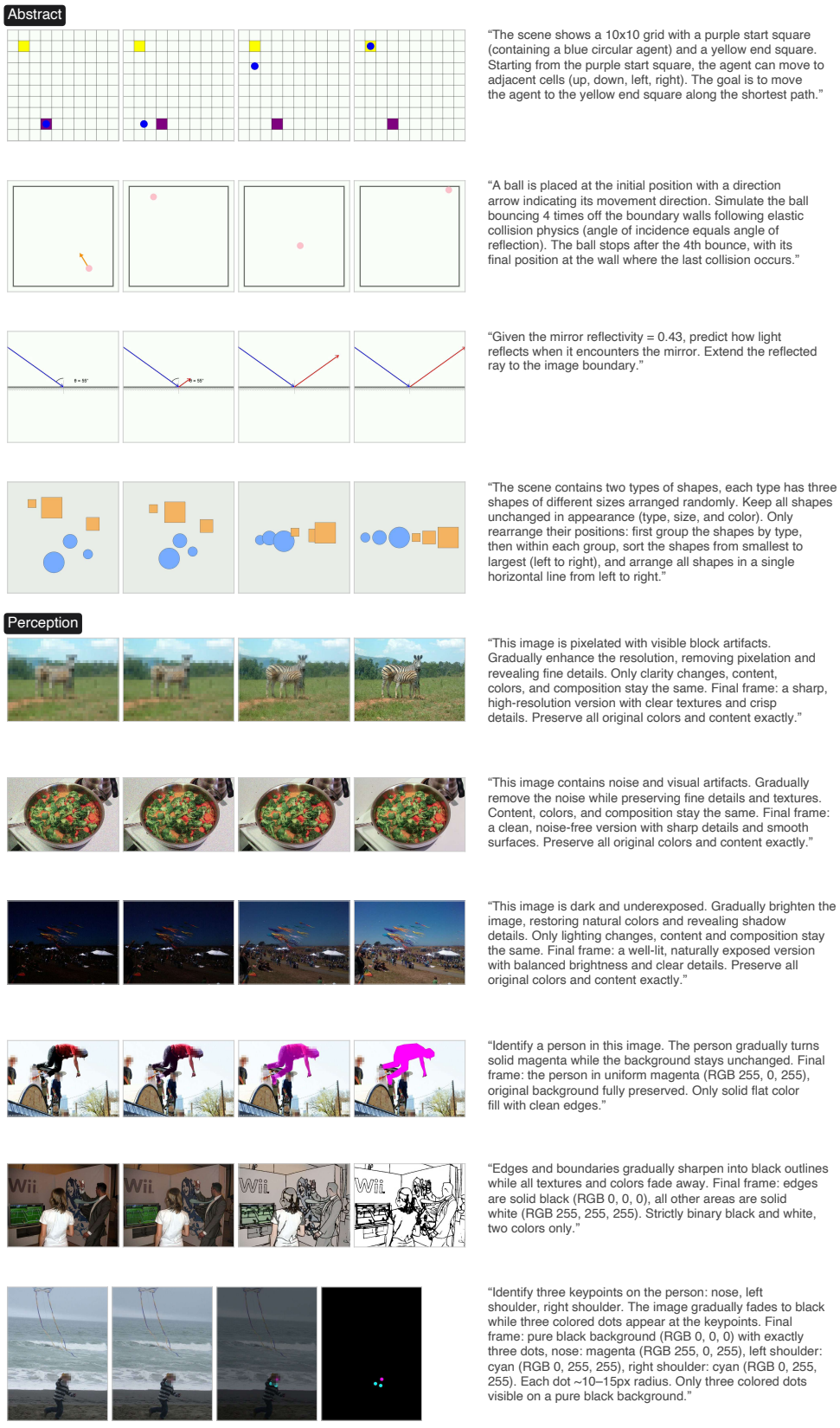

Figure 7: Training samples: abstract reasoning and perception.

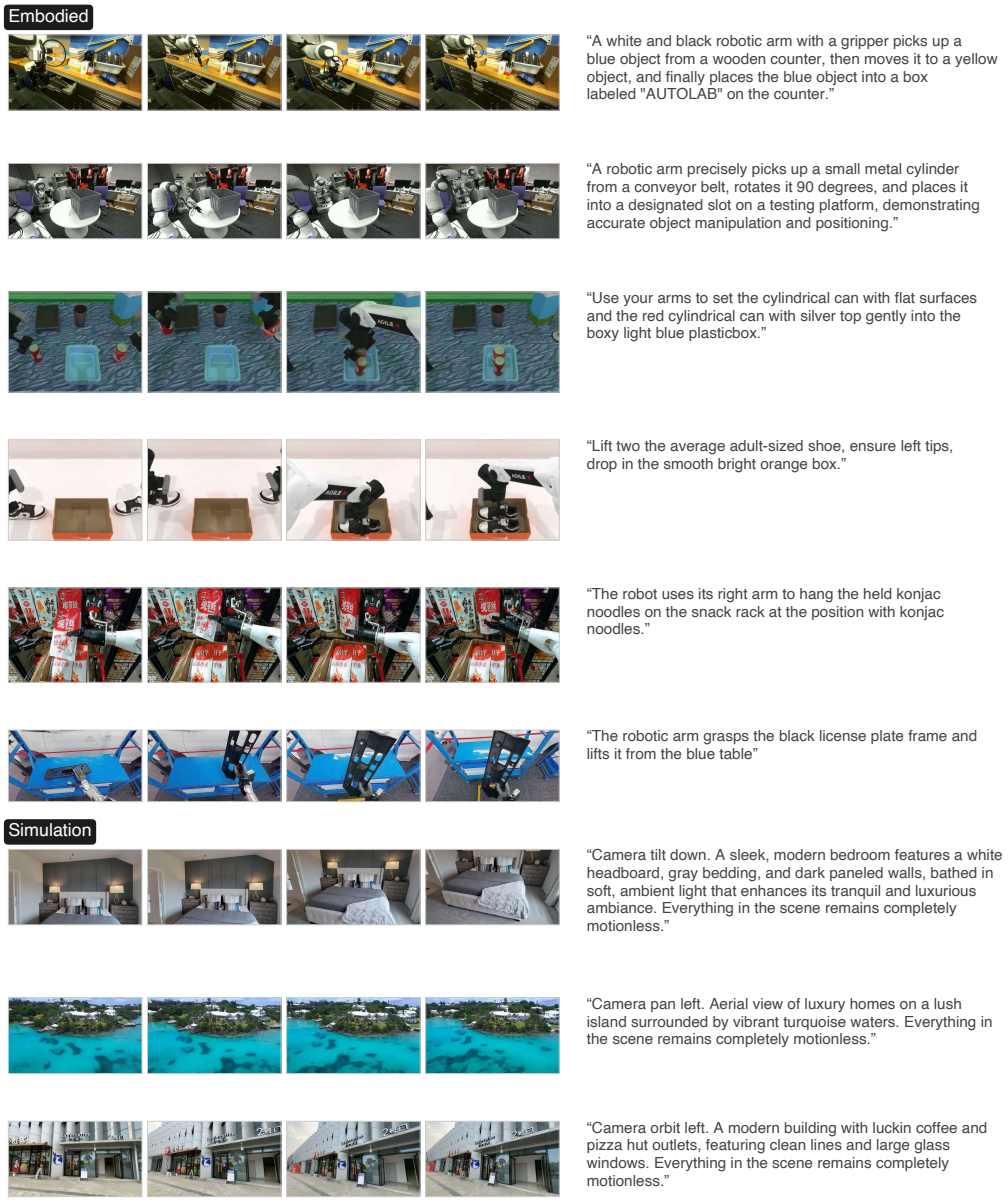

Figure 8: Training samples: embodied manipulation and simulation.

E BENCHMARK SCORING DETAILS

E.1 ABSTRACT REASONING

Equation 1 weights VBVR-Bench by $1/2$ and each of the other four benchmarks by $1/8$; all five scores lie in $[0, 100]$. VBVR-Bench uses its 500 instances (50 ID and 50 OOD task types, five instances each), Original prompts, and the task-specific rule-based evaluator. Its score is $(S_{\mathrm{ID}} + S_{\mathrm{OOD}})/2$. The ID/OOD names follow the benchmark's original task partition.

VIPER uses its full set of 309 items over 16 tasks and six reasoning domains and reports POC@1, requiring both a correct outcome and a consistent process, judged with GPT-5. Gen-ViRe reports its official overall Avg across six reasoning dimensions using its released evaluation pipeline with Gemini 2.5 Pro. V-ReasonBench reports pass@1 over its 13 tasks: instances are pooled within each of four dimensions, then the dimension scores are averaged. It combines mask-based, grid-based, and VLM-based evaluators. VideoThinkBench uses its 500 vision-centric items: eyeballing (210), maze (50), ARC (140), and visual puzzle (100). We average the four family accuracies using last-frame (LF) evaluation. Multi-frame scores remain auxiliary. Table 8 reports the five component scores and their domain average.

E.2 PERCEPTION PER-TASK METRICS

Every metric m_t compares a candidate frame against the ground-truth target image at a common resolution and returns a score in $[0, 1]$; Equation 2 takes the maximum over the eight sampled frames, and the six per-task means are rescaled to $[0, 100]$ and averaged.

Segmentation. The predicted mask is extracted in HSV space as the magenta region ($H \in [140, 170]$ on OpenCV's 0–180 scale, $S > 100$, $V > 100$), cleaned by a 3×3 elliptical morphological opening with connected components under 30 px discarded; the ground-truth mask is extracted identically from the target image. The score is $\mathrm{IoU} = |M_{\text{pred}} \cap M_{\text{gt}}| \,/\, |M_{\text{pred}} \cup M_{\text{gt}}|$, defined as 0 when the union is empty.

Edge detection. A sanity gate first measures the fraction of near-white pixels ($V > 180$, $S < 60$); frames below a white fraction of 0.40 score 0, which blocks a photograph from being rewarded by chance overlap with dilated edges (ground-truth edge maps are 78–97% white, photographs at most 35%). Predicted edges are the pixels with $V < 30$, thinned by non-maximum suppression with components under 30 px dropped; ground-truth edges are the target pixels with $V < 128$. With tolerance $\tau = \max\big(1, \mathrm{round}(0.015\sqrt{h^2 + w^2})\big)$, both binary maps are dilated by an elliptical element of radius τ, and the score is the F-measure of $\mathrm{Prec} = |E_{\text{pred}} \cap E_{\text{gt}}^{(\tau)}|/|E_{\text{pred}}|$ and $\mathrm{Rec} = |E_{\text{gt}} \cap E_{\text{pred}}^{(\tau)}|/|E_{\text{gt}}|$; two empty maps score 1, exactly one empty map scores 0.

Keypoint localization. The nose is the centroid of the largest magenta blob and the shoulders are the centroids of the two largest cyan blobs ($H \in [80, 100]$), extracted with the same thresholds and cleanup as segmentation. A predicted point $\hat{p}$ scores $s = \max\big(0,\, 1 - \|\hat{p} - p\|_2/(0.1D)\big)$ against its ground-truth point p, where D is the image diagonal. Because both shoulders share one color, the two predicted blobs are assigned to the (left, right) ground-truth pair by whichever of the two matchings scores higher; a missing blob scores 0. The frame score is the mean of the three point scores.

Restoration. The three restoration tasks share one score, which combines reference-based perceptual fidelity (Zhang et al., 2018) with no-reference image quality (Ke et al., 2021) under a brightness-matching factor:

$$m_{\text{restore}} \;=\; \sqrt{\big(1 - \mathrm{LPIPS}\big) \cdot \tfrac{\mathrm{MUSIQ}}{100}} \;\cdot\; \mathrm{bmatch}, \qquad \mathrm{bmatch} = \max\Big(0,\, 1 - \tfrac{|\bar{V}_{\text{pred}} - \bar{V}_{\text{gt}}|}{255}\Big), \quad (4)$$

where $\bar{V}$ is the mean HSV value channel. The geometric mean makes it impossible to score well on perceptual quality alone, and bmatch blocks the "over-brighten to daylight" failure mode that no-reference quality metrics actively reward.

E.3 EMBODIED SCORING

We use RBench's automatic Task Completion protocol as released (Deng et al., 2026). Each generated video is converted into a temporal grid of uniformly sampled frames and judged by GPT-5, the benchmark's closed-source MLLM evaluator, with a VQA-style checklist. Two components are scored on a five-point scale and averaged: physical–semantic plausibility and non-contact attachment or incorrect grasps; and task-adherence consistency, which checks task responsiveness (the goal state is reached without premature interruption) and key actions (the prompted actions occur and in the prompted order). RBench reports a Spearman correlation of 0.96 between its overall automatic score and human preference over ten models; we do not repeat the human study.

E.4 SIMULATION SCORING

We use the WorldScore protocol as released (Duan et al., 2025). Three measure controllability. *Camera control* estimates per-frame camera poses with DROID-SLAM (Teed & Deng, 2021). *Object control* is the detection rate of the objects named in the next-scene prompt, using an open-set detector (Liu et al., 2024). *Content alignment* is the CLIPScore (Hessel et al., 2021) between the generated scene and the full next-scene prompt. Four measure quality. *3D consistency* is the dense reprojection error of co-visible pixels after DROID-SLAM bundle adjustment; *photometric consistency* is the forward–backward optical-flow end-point error between consecutive frames; *style consistency* is the Gram-matrix distance between the first and last frame; and *subjective quality* averages CLIP-IQA+ and CLIP Aesthetic, the combination WorldScore found to agree best with human raters.

E.5 SIMULATION AGGREGATION

Because `content_alignment` is defined only for multi-segment instances, it has 700 samples while the other six aspects have 1,000, so aspects are paired separately during aggregation. The calibrated protocol is

$$
\begin{aligned}
A_{\text{cam}} &= \frac{100}{1+\exp\!\big(-(c-T_{\text{base}})/\tau\big)}, \qquad \tau = 2.5, \\
\text{Ability} &= \text{mean}\big(A_{\text{cam}},\ \text{object control},\ \text{content alignment},\ \text{subjective quality}\big), \\
\text{Consistency} &= \text{mean}\big(\text{3D},\ \text{photometric},\ \text{style}\big), \\
Simulation - Calibrated &= 0.7 \cdot \text{Ability} + 0.3 \cdot \text{Consistency}, \qquad (5)
\end{aligned}
$$

where T_{base} is the base model's own camera-control score (32.45 for Wan2.2 and 24.58 for LTX-2.3), so that $A_{\text{cam}} = 50$ for the base model by construction.

F VISUAL DOMAIN TRANSFORMATION AND EVALUATION

Each first-frame transformation starts from an image-specific edit instruction. Nano Banana 2 (`gemini-3.1-flash-image`) produces a visually different rendition, and a VLM checks that the answer-bearing facts remain intact, retrying edits that alter the task. The same transformed inputs are shared by the baseline and all adapters. Depending on the task, the edit preserves geometric layout, letters and digits, grid structure, colors, or object relations. Geometric-construction items additionally use aligned ink-mask overlap to check preservation of the diagram (IoU threshold 0.30).

The transformed sets cover 309 VIPER inputs, 72 Gen-ViRe inputs, six VLM-assessed V-ReasonBench tasks (193 inputs), and the eyeballing and visual-puzzle families of VideoThinkBench (310 inputs). Eighteen VIPER inputs already show real-world scenes or text-heavy pages that need no editing, so their original images are used. The V-ReasonBench subset comprises math, sudoku, code, temperature, shape fitting, and color connection. Shifted VideoThinkBench uses a VLM judge, Gemini 3.8 Flash (`gemini-3.8-flash`) (Google DeepMind, 2026a), and reports pooled accuracy across its 310 items. Its scores are compared between models under this shared shifted protocol, separately from the original rule-based evaluation. The five shifted benchmark scores are combined with the weights of Equation 1, as the original scores are.

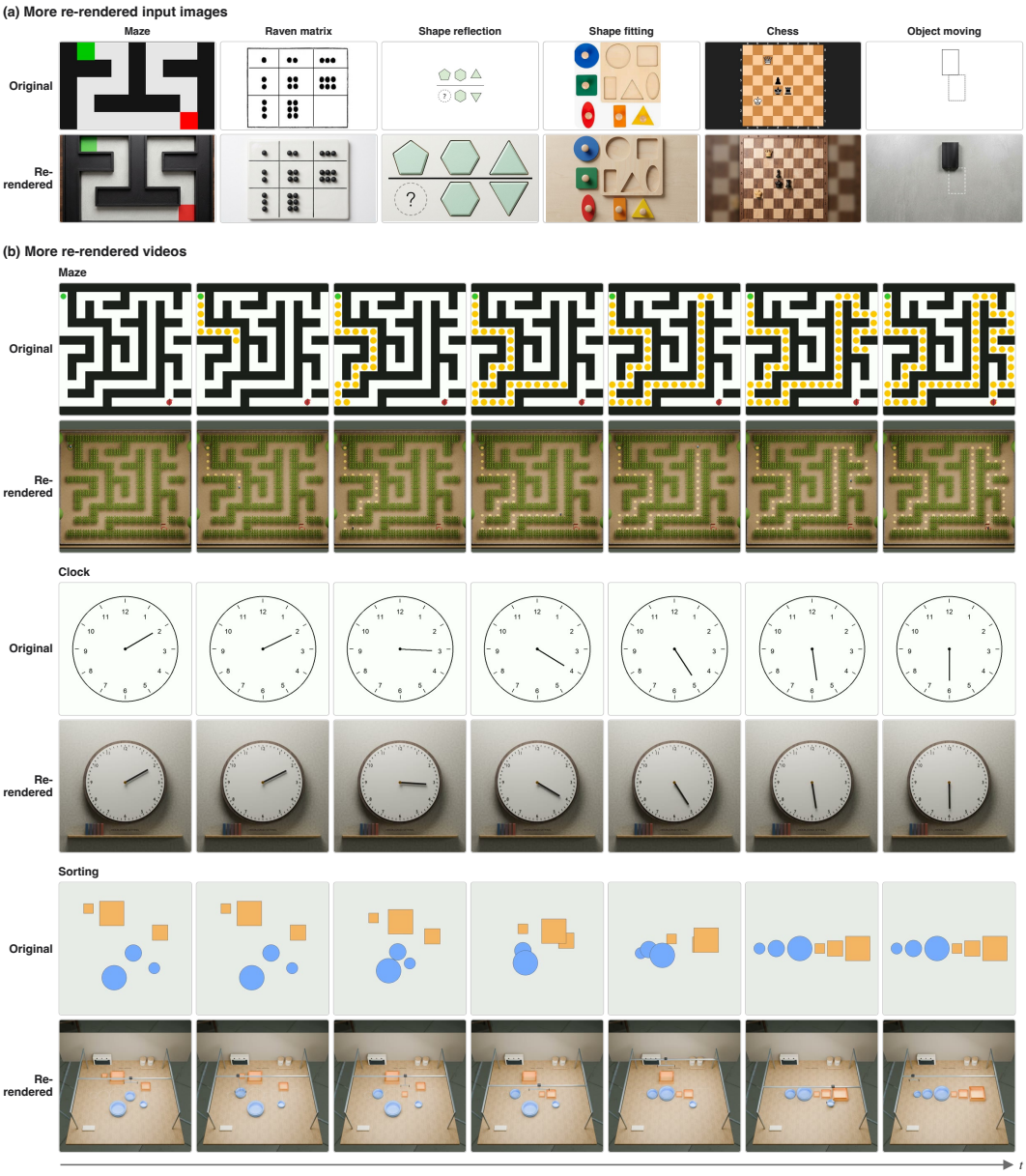

Figure 9: More visual domain shifts. **(a)** Original and re-rendered input images from the image-only benchmarks. **(b)** Seven frames of three VBVR-Bench videos and their Blender re-renderings.

For VBVR-Bench, the procedural generator maps the task state and its valid solution to corresponding Blender objects and rendered scenes. Rule-based evaluation must recover the task-relevant state from the transformed rendering. Each of its 500 items is re-rendered as a corresponding Blender video; as for the original split, the VBVR score is the mean of the ID and OOD splits. Shifted V-ReasonBench reports item-weighted pass@1 over the 193 inputs, counting a final frame with no readable answer as a failure.

Figure 9 shows further examples of both routes.

G REASONING AND INSTRUCTION-FOLLOWING SUBSETS

The abstract-reasoning corpus asks for two distinguishable things: multi-step inference toward a goal, and faithful execution of an explicitly given instruction. To ask which component drives generalization, we partition the VBVR training data at the level of task types by a single criterion: whether the answer the video must produce is fully specified in the prompt. If it is ("move the object left off-screen, then return along the same path"), the task is *instruction-following*; if the target state has to be inferred ("select the option that continues the sequence and circle it"), it is *reasoning*. Two annotators labeled all 100 VBVR training task types independently and reconciled disagreements, giving 47 reasoning and 53 instruction-following types (on the benchmark, 31/19 of the ID and 13/37 of the OOD task types). We then sample an equal number of clips from each subset and train under an identical budget.

Both subsets improve over the baseline on their own (Figure 10a). On the three application domains, the instruction-following subset leads on every one, with Reasoning FT shown alongside them. Instruction following therefore transfers more strongly to the application domains. More importantly, the reasoning subset, which contains no instruction-following task types, still improves over the baseline on all three, so reasoning is itself a capability that generalizes across domains. Splitting the abstract benchmark by the same typology (Figure 10b) shows each subset leading on its own task type on both splits: the two subsets train largely independent capabilities, each generalizing within its own type. Since both subsets share the same synthetic visual distribution yet generalize differently, the task structure of the data is doing the work.

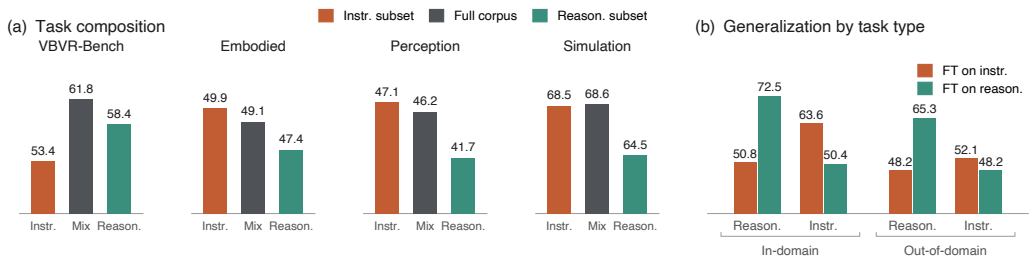

Figure 10: Reasoning vs. instruction-following VBVR subsets on Wan2.2. **(a)** Task composition: LoRA fine-tuning on the instruction-following subset, the full corpus (Mix), and the reasoning subset; the abstract axis is VBVR-Bench. **(b)** VBVR-Bench split by task type.

G.1 PARAMETER-LEVEL ANALYSIS OF THE LORA UPDATES

For every LoRA checkpoint, we expand the update $\Delta W = BA$ of each adapted matrix and measure the share of the total squared Frobenius norm $\sum \|\Delta W\|^2$ by module type and by depth, with each checkpoint normalized to 100%; Figure 12 plots the per-block depth profiles.

The two subsets leave distinct footprints in the weights. On Wan2.2's high-noise expert at 50K, the instruction-following LoRA allocates more of its squared update norm to text cross-attention (23.0% vs. 17.6%) and to the early third of the blocks (44.2% vs. 37.8%), whereas the reasoning LoRA allocates more to the FFNs (52.5% vs. 47.8%) and to the late blocks (31.8% vs. 25.8%). The two subsets differ in *where* they write, not in how much: instruction following reweights text conditioning in early blocks, reasoning reshapes the later computation that carries cross-frame structure (Figure 11). Because the two components serve different targets, the practical implication is to combine rather than choose between them.

We have four further observations. **(i)** On Wan2.2, the instruction-following LoRA consistently allocates more of its update to text cross-attention and to the early third of the blocks than the reasoning LoRA, on both experts and at every scale, while the reasoning LoRA allocates more to the FFNs and to the late blocks. The contrast is sharpest on the high-noise expert, the expert that decides layout and semantics. **(ii)** On LTX-2.3 the module-level contrast moves to video self-attention, the module that carries cross-frame structure (55.7% vs. 51.2% at 50K), while the depth-level contrast has the same sign as on Wan2.2 and is larger: 64.6% vs. 57.4% of the update lands in the last third

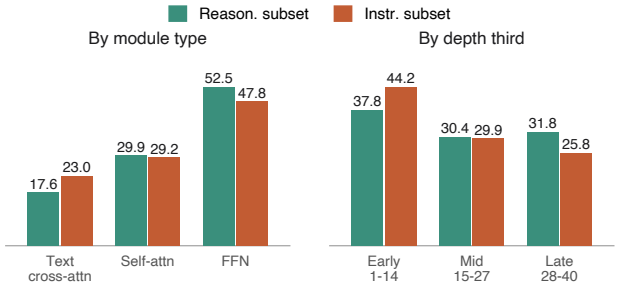

Figure 11: Where the two subsets write their LoRA updates: share of the total squared update norm by module type and by depth third (Wan2.2 high-noise expert, 50K).

of the blocks, 6.8% vs. 10.9% in the first. **(iii)** The full-corpus LoRA sits between the two subsets on nearly every axis, as expected for a mixture. **(iv)** Total update magnitudes are comparable between the two subsets at matched scale ($\sqrt{\sum \|\Delta W\|^2}$ of 27.4 vs. 27.1 on the Wan2.2 high-noise expert at 50K), so the differences above are differences of placement, not of strength. All four patterns are stable across 50K, 200K, and 500K.

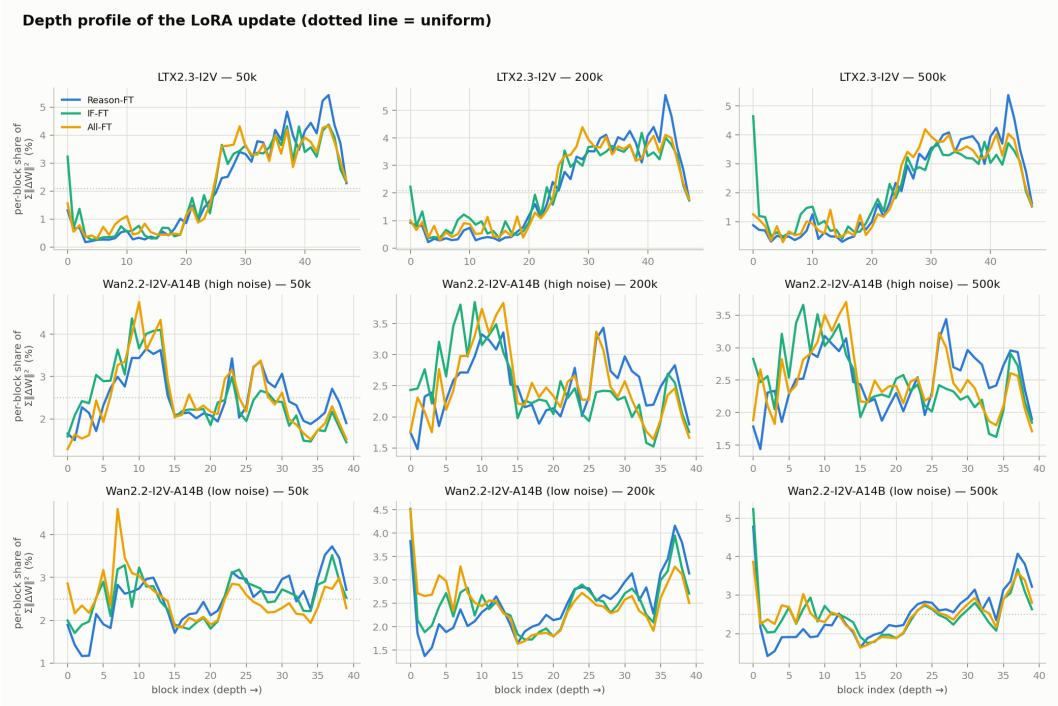

Figure 12: Per-block share of the squared LoRA update norm as a function of depth, for both backbones and three training scales; the dotted line marks a uniform allocation.

Table 7: Absolute scores underlying Figure 4, shown as Wan2.2 / LTX-2.3. Reasoning is the weighted five-benchmark score of Equation 1. Simulation uses Simulation-Calibrated; diagonal cells are bold.

Training data	Reasoning	Embodiment	Perception	Simulation
Baseline	27.2 / 25.7	40.7 / 35.2	32.5 / 28.8	62.0 / 58.7
Reasoning FT	**44.6 / 38.8**	49.1 / 39.3	46.2 / 36.5	68.6 / 66.0
Embodiment FT	27.6 / 27.3	**48.8 / 40.9**	38.9 / 33.8	67.1 / 52.2
Perception FT	31.0 / 27.1	38.2 / 22.1	**78.7 / 68.7**	61.4 / 55.6
Simulation FT	24.1 / 23.5	47.3 / 35.1	34.9 / 30.6	**69.4 / 70.4**

Table 8: Abstract reasoning across five benchmarks. VBVR is the mean of Original-prompt ID and OOD scores. Average weights VBVR by 0.5 and each other benchmark by 0.125 (Equation 1); all scores are on $[0, 100]$.

Training data	VBVR	VIPER	Gen-ViRe	V-Reason	VideoThink	Average
Wan2.2						
Baseline	37.1	3.0	29.1	4.2	32.9	27.2
Reasoning FT (10K)	58.7	6.0	43.0	3.9	47.2	41.8
Reasoning FT	61.8	8.0	45.1	10.4	46.1	44.6
high-noise only	59.8	5.0	45.3	11.6	45.1	43.3
low-noise only	34.6	1.0	33.4	2.9	35.1	26.3
Reasoning FT (200K)	63.7	5.0	46.2	16.8	47.6	46.3
Reasoning FT (500K)	62.7	10.0	44.4	16.7	47.2	46.1
Reasoning full FT	60.0	5.0	42.7	15.5	46.2	43.7
Perception FT	42.3	7.0	35.9	2.5	33.3	31.0
Embodiment FT	36.5	3.0	32.3	4.5	35.1	27.6
Simulation FT	31.2	2.0	28.3	4.5	33.2	24.1
LTX-2.3						
Baseline	35.1	1.0	29.0	5.8	29.4	25.7
Reasoning FT	53.0	4.0	39.7	7.5	46.7	38.8
Perception FT	36.5	1.0	34.7	3.7	31.1	27.1
Embodiment FT	36.3	4.0	33.0	5.4	30.3	27.3
Simulation FT	33.7	1.0	23.3	5.8	23.1	23.5

H FULL RESULT TABLES

Table 7 lists the absolute benchmark scores behind the relative cross-domain matrix of Figure 4. Tables 9–13 then report, for every fine-tuned model on both backbones, the per-task or per-aspect sub-scores of the original benchmarks; the abstract diagnostic table is specifically VBVR-Bench, while Table 8 reports the five-benchmark domain score; all scores are on a $[0, 100]$ scale. Tables 14–17 give the sub-scores of the other four reasoning benchmarks, and Tables 18–21 give the per-benchmark visual-shift scores. Rows marked high-noise only and low-noise only load the Reasoning FT LoRA into a single Wan2.2 expert (§5.3); Reasoning full FT is the full fine-tuning run of §5.1. Table 22 reports VBench quality dimensions for the base model and the reasoning-tuned model on both backbones.

Figure 13 gives three further qualitative cases per application domain, in the same format as Figure 6.

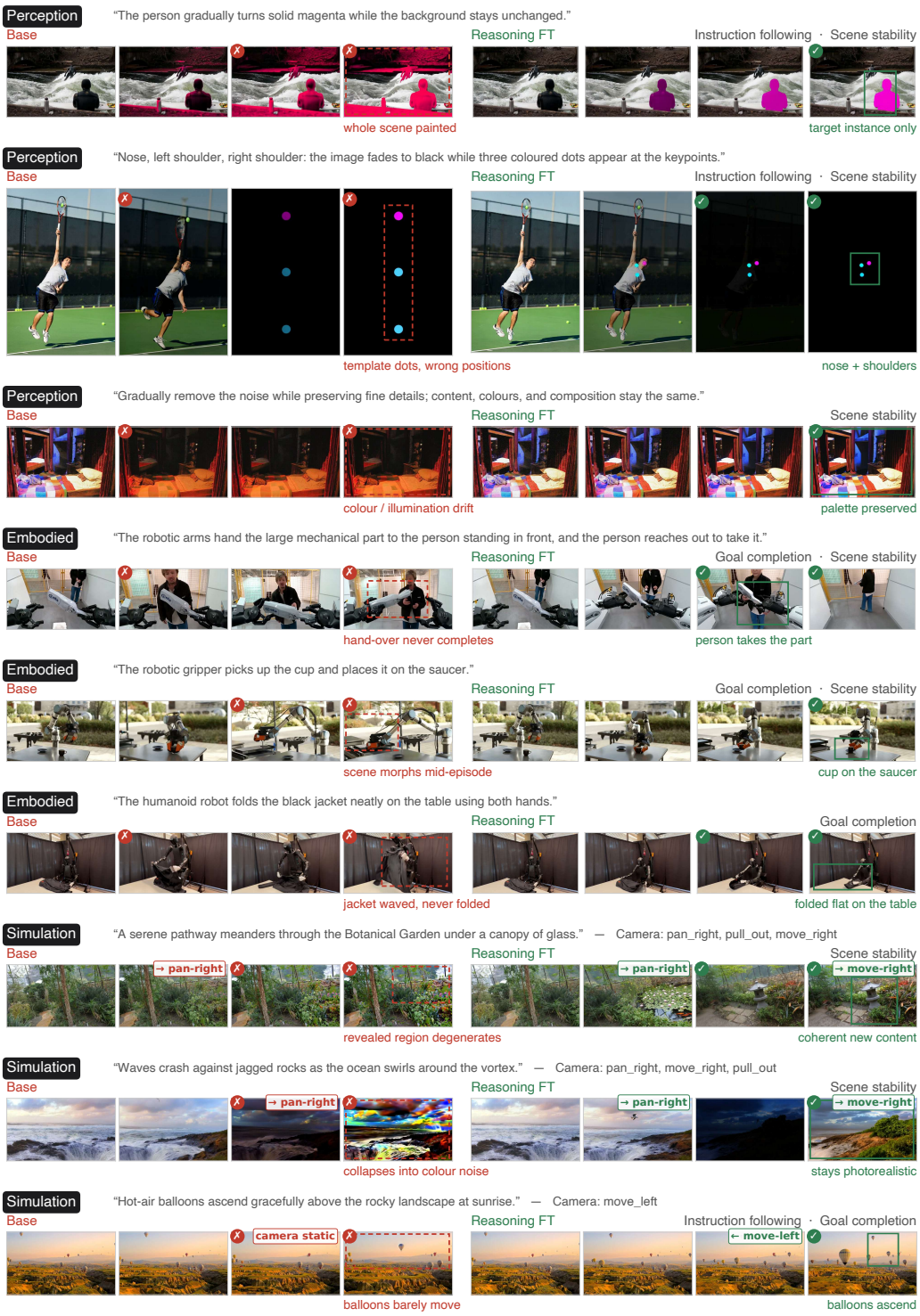

Figure 13: Additional qualitative cases, three per application domain.

Table 9: VBVR-Bench diagnostic results. Overall is the mean of the Original-prompt ID and OOD splits; the remaining columns are the ID/OOD × reasoning/instruction cells and the two task-type marginals. Rows under a backbone name are fine-tuned from that backbone and labeled by training data. For third-party models we take the ID/OOD scores reported by VBVR-Bench, which does not release per-task scores, so the reasoning/instruction breakdown is unavailable (–).

Model / data	**Overall**	**ID**	**OOD**	**ID**		**OOD**		**All**	
				Reas.	Instr.	Reas.	Instr.	Reas.	Instr.
Wan2.2-I2V-A14B									
Baseline	37.1	41.2	32.9	42.7	38.7	46.4	28.2	43.8	31.7
Reasoning FT (10K)	58.7	63.4	53.9	63.6	63.1	60.7	51.5	62.8	55.4
Reasoning FT	61.8	67.8	55.9	68.7	66.2	59.0	54.8	65.9	58.6
high-noise only	59.8	64.6	55.0	64.8	64.1	58.9	53.7	63.1	57.2
low-noise only	34.6	40.1	29.0	41.6	37.7	41.6	24.5	41.6	29.0
Reasoning FT (200K)	63.7	73.4	53.9	74.6	71.5	61.5	51.2	70.8	58.1
Reasoning FT (500K)	62.7	72.9	52.5	72.0	74.5	59.7	50.0	68.4	58.3
reasoning subset	58.4	64.1	52.6	72.5	50.4	65.3	48.2	70.4	48.9
reasoning subset 200K	56.5	63.2	49.8	72.8	47.4	63.9	44.9	70.2	45.7
reasoning subset 500K	53.7	63.1	44.3	73.1	46.9	59.3	39.0	69.0	41.7
instruction subset	53.4	55.7	51.1	50.8	63.6	48.2	52.1	50.0	56.0
instruction subset 200K	54.1	60.2	48.1	53.0	71.8	50.1	47.3	52.2	55.6
instruction subset 500K	54.6	60.8	48.3	51.9	75.4	49.6	47.9	51.2	57.2
Reasoning full FT	60.0	71.1	48.9	71.8	70.0	56.1	46.3	67.2	54.3
Perception FT (10K)	40.2	45.0	35.4	45.6	44.2	40.0	33.7	43.9	37.3
Perception FT	42.3	45.6	39.0	43.9	48.4	43.8	37.3	43.9	41.0
Perception, full set	42.5	46.9	38.2	46.4	47.7	43.7	36.2	45.6	40.1
Embodiment FT	36.5	42.3	30.7	42.8	41.5	45.6	25.5	43.7	30.9
Embodiment FT (100K)	36.9	42.0	31.9	42.6	41.0	43.9	27.6	43.0	32.2
Simulation FT	31.2	37.2	25.3	37.5	36.7	37.7	20.9	37.6	26.3
LTX-2.3-I2V									
Baseline	35.1	40.4	29.8	41.3	39.1	37.6	27.0	40.2	31.1
Reasoning FT	53.0	58.6	47.5	61.2	54.4	62.4	42.3	61.5	46.4
Reasoning FT (200K)	55.0	62.4	47.6	65.5	57.4	56.8	44.3	62.9	48.8
Reasoning FT (500K)	53.4	61.9	44.9	64.3	57.9	51.6	42.5	60.5	47.7
reasoning subset	51.5	58.4	44.7	65.5	46.9	53.7	41.5	62.0	43.3
reasoning subset 200K	53.8	61.8	45.7	69.5	49.3	60.8	40.5	66.9	43.4
reasoning subset 500K	52.4	60.4	44.5	68.0	47.9	60.6	38.9	65.8	42.0
instruction subset	48.1	50.9	45.3	47.4	56.5	48.6	44.2	47.7	48.4
instruction subset 200K	50.4	53.2	47.6	45.5	65.6	54.1	45.3	48.1	52.2
instruction subset 500K	46.0	49.9	42.2	42.0	62.7	47.9	40.2	43.8	47.8
Perception FT (10K)	37.1	41.0	33.3	41.5	40.2	44.4	29.4	42.4	33.0
Perception FT	36.5	39.8	33.2	38.8	41.5	43.1	29.8	40.1	33.8
Perception, full set	35.5	39.4	31.5	38.7	40.4	43.7	27.3	40.2	31.8
Embodiment FT	36.3	41.7	31.0	43.2	39.1	44.3	26.3	43.6	30.7
Embodiment FT (100K)	33.8	39.0	28.6	41.3	35.1	41.6	24.0	41.4	27.8
Simulation FT	33.7	39.1	28.3	39.9	37.8	35.5	25.8	38.6	29.9
Commercial models									
Sora 2	54.6	56.9	52.2	–	–	–	–	–	–
Veo 3.1	48.0	53.1	42.9	–	–	–	–	–	–
Kling 2.6	36.9	40.8	33.0	–	–	–	–	–	–
Other open-source models									
LTX-2	31.3	32.9	29.7	–	–	–	–	–	–
CogVideoX1.5-5B	27.3	28.3	26.2	–	–	–	–	–	–
HunyuanVideo-I2V	27.3	28.0	26.5	–	–	–	–	–	–
Wan2.2-TI2V-5B	32.3	39.2	25.4	–	–	–	–	–	–

Table 10: Perception: per-task scores and their mean. Sora 2 refuses the keypoint task under its content policy; its mean is over the remaining five tasks.

Model / data	Seg.	Edge	SR	Low-light	Denoise	Keypoint	Mean
Wan2.2-I2V-A14B							
Baseline	36.9	4.7	31.9	54.9	59.0	7.4	32.5
Reasoning FT (10K)	65.7	11.8	39.4	53.1	66.4	23.8	43.4
Reasoning FT	70.7	24.4	28.6	58.5	68.0	27.0	46.2
high-noise only	71.1	24.6	28.7	58.3	68.3	27.2	46.4
low-noise only	45.2	14.3	36.9	45.4	61.0	12.7	35.9
Reasoning FT (200K)	70.8	15.5	28.2	63.2	67.0	25.9	45.1
reasoning subset	66.5	4.6	30.7	60.1	65.2	22.9	41.7
reasoning subset 200K	63.8	9.8	29.7	61.5	64.5	14.0	40.6
instruction subset	68.6	20.8	33.1	59.3	67.8	33.2	47.1
instruction subset 200K	69.8	18.8	35.9	60.9	67.5	34.1	47.8
Reasoning full FT	70.4	15.4	38.7	61.0	62.0	19.9	44.6
Perception FT (10K)	76.4	93.9	63.3	68.1	71.9	83.0	76.1
Perception FT	81.5	97.7	65.0	69.8	71.7	86.7	78.7
Perception, full set	82.5	98.0	65.2	70.2	71.3	87.6	79.2
Embodiment FT	55.5	5.3	43.5	55.2	62.7	11.4	38.9
Simulation FT	41.0	11.1	43.5	47.7	57.8	8.4	34.9
ReasonGraft (3:7)	83.5	97.1	65.2	69.8	72.2	85.4	78.9
ReasonGraft (5:5)	83.0	92.5	64.6	69.4	72.1	83.2	77.5
ReasonGraft (7:3)	81.1	85.1	62.0	64.2	70.9	74.2	72.9
Mixed-data FT (3:7)	82.2	95.3	63.9	70.4	70.8	87.3	78.3
LTX-2.3-I2V							
Baseline	37.8	28.0	26.8	20.4	50.2	9.4	28.8
Reasoning FT	65.8	24.9	27.3	40.7	57.8	2.3	36.5
Reasoning FT (200K)	62.5	23.4	27.2	34.3	58.0	1.1	34.4
Reasoning FT (500K)	63.0	10.3	27.2	51.4	59.6	2.1	35.6
reasoning subset	62.3	20.0	27.1	40.4	53.9	6.3	35.0
reasoning subset 200K	57.9	14.7	27.2	42.5	53.8	5.2	33.6
reasoning subset 500K	59.4	6.5	27.4	42.7	52.0	5.8	32.3
instruction subset	63.9	38.1	26.9	35.0	57.8	7.8	38.2
instruction subset 200K	62.4	32.0	26.9	30.6	59.5	4.8	36.0
instruction subset 500K	62.0	15.6	27.1	32.8	62.8	4.3	34.1
Perception FT (10K)	74.1	94.7	47.8	61.7	63.3	55.9	66.3
Perception FT	79.1	96.7	38.4	60.0	57.5	80.4	68.7
Perception, full set	84.8	97.3	35.9	59.4	55.0	82.5	69.2
Embodiment FT	58.0	23.3	27.5	37.2	48.4	8.4	33.8
Embodiment FT (100K)	50.7	13.2	27.5	36.8	48.7	6.7	30.6
Simulation FT	42.3	18.4	25.6	41.1	47.4	9.1	30.6
ReasonGraft (3:7)	74.1	93.7	49.7	63.6	64.0	69.7	69.1
Commercial models							
Sora 2	65.5	43.4	46.1	42.6	54.1	–	50.3
Veo 3.1	59.0	28.6	52.5	55.7	63.9	48.3	51.4
Kling 2.6	12.3	9.9	40.2	45.9	59.7	36.9	34.1
Other open-source models							
LTX-2	51.0	52.7	26.0	36.8	46.4	10.4	37.2
CogVideoX1.5-5B	8.2	0.0	27.6	27.5	28.4	10.4	17.0
HunyuanVideo-I2V	37.1	3.4	27.3	48.7	48.3	12.9	29.6
Wan2.2-TI2V-5B	23.6	10.2	27.2	39.4	46.9	5.7	25.5

Table 11: Embodied manipulation (RBench task-completion, five task families) and their mean.

Model / data	Common	Long-horizon	Multi-entity	Spatial	Visual reas.	Mean
Wan2.2-I2V-A14B						
Baseline	40.3	50.2	32.5	42.8	37.6	40.7
Reasoning FT (10K)	52.4	57.3	49.1	54.1	39.9	50.6
Reasoning FT	49.0	60.8	45.4	51.2	39.3	49.1
high-noise only	52.9	49.2	50.7	49.6	39.4	48.3
low-noise only	50.0	49.2	43.5	47.2	37.5	45.5
Reasoning FT (200K)	46.7	62.5	34.9	53.5	42.5	48.0
Reasoning FT (500K)	40.4	56.1	34.7	49.8	41.9	44.6
reasoning subset	44.9	54.0	46.0	47.5	44.8	47.4
instruction subset	48.8	62.8	41.2	55.8	41.1	49.9
Reasoning full FT	31.7	35.9	22.6	45.0	25.1	32.0
Perception FT (10K)	37.2	53.3	36.6	39.4	33.4	40.0
Perception FT	36.9	51.0	37.8	35.4	29.9	38.2
Perception, full set	32.0	48.3	26.3	31.2	23.6	32.3
Embodiment FT	45.9	53.6	44.7	56.6	43.2	48.8
Simulation FT	58.1	48.4	47.2	44.9	37.9	47.3
ReasonGraft (3:7)	55.6	56.2	48.1	57.1	44.7	52.3
ReasonGraft (5:5)	56.5	59.8	44.3	53.5	46.4	52.1
ReasonGraft (7:3)	52.1	58.2	45.4	48.0	39.7	48.7
Mixed-data FT (3:7)	43.7	42.5	42.8	44.2	31.4	40.9
LTX-2.3-I2V						
Baseline	38.8	48.3	29.0	38.6	21.6	35.2
Reasoning FT	30.8	50.0	46.3	43.6	25.9	39.3
Reasoning FT (200K)	30.2	49.8	33.4	41.4	23.8	35.7
reasoning subset	41.8	50.1	39.2	45.6	22.0	39.7
reasoning subset 200K	34.6	47.4	28.5	38.4	22.0	34.2
instruction subset	39.1	55.4	37.4	49.5	31.0	42.5
instruction subset 200K	32.2	46.5	33.9	49.7	23.1	37.1
Perception FT	21.4	32.8	17.9	23.6	14.6	22.1
Embodiment FT	42.8	39.6	42.8	43.1	36.0	40.9
Simulation FT	36.7	48.1	31.8	36.4	22.5	35.1
ReasonGraft (3:7)	47.2	54.8	48.3	48.3	45.1	48.8
Commercial models						
Sora 2	20.8	25.5	18.6	26.8	11.5	20.6
Veo 3.1	40.4	43.7	43.6	59.5	45.3	46.5
Kling 2.6	52.9	53.0	36.4	59.8	35.8	47.6
Other open-source models						
LTX-2	28.4	38.6	23.3	30.4	16.4	27.4
CogVideoX1.5-5B	11.6	21.2	9.8	11.2	7.9	12.3
HunyuanVideo-I2V	44.2	43.8	31.2	31.6	36.4	37.4
Wan2.2-TI2V-5B	33.1	31.8	14.2	31.3	23.4	26.8

Table 12: World simulation (WorldScore-Static): the seven aspects and their absolute mean. Sora 2, Veo 3.1, and Kling 2.6 were not evaluated on this benchmark.

Model / data	**Camera**	**Object**	**Content**	**Subj.**	**3D**	**Photom.**	**Style**	**Absolute**
Wan2.2-I2V-A14B								
Baseline	32.5	42.0	58.9	59.9	78.6	85.6	86.7	63.5
Reasoning FT (10K)	35.6	42.9	61.3	59.4	79.3	86.4	86.9	64.5
Reasoning FT	36.8	42.6	59.6	59.6	80.3	86.6	87.6	64.7
high-noise only	35.6	42.4	60.4	59.1	79.5	86.5	86.9	64.3
low-noise only	32.9	42.5	59.3	60.5	78.7	85.5	87.8	63.9
Reasoning FT (200K)	34.9	43.6	56.4	60.0	82.3	89.4	89.6	65.2
reasoning subset	33.2	43.9	52.0	61.3	85.1	92.2	91.6	65.6
reasoning subset 200K	34.1	43.6	55.0	61.7	84.7	90.1	90.3	65.6
instruction subset	36.4	42.1	59.8	60.5	80.4	87.8	88.2	65.0
instruction subset 200K	36.2	42.6	58.8	59.6	80.0	87.1	87.5	64.5
Reasoning full FT	31.8	40.4	45.4	56.6	80.6	89.9	88.9	62.0
Perception FT	31.5	42.9	53.5	59.5	87.0	94.6	88.5	65.4
Embodiment FT	35.1	42.3	60.4	59.6	81.0	88.7	87.0	64.9
Simulation FT	42.9	41.0	58.2	56.9	74.4	86.1	87.9	63.9
ReasonGraft (3:7)	38.7	42.0	60.9	59.3	77.7	86.1	87.3	64.6
ReasonGraft (5:5)	40.2	41.8	61.1	58.8	76.7	85.9	87.3	64.5
ReasonGraft (7:3)	41.4	41.2	59.1	58.2	75.9	86.1	88.0	64.3
Mixed-data FT (3:7)	43.7	40.9	55.5	58.4	75.2	87.3	88.7	64.2
LTX-2.3-I2V								
Baseline	24.6	61.0	73.7	46.7	57.5	75.2	49.2	55.4
Reasoning FT	28.3	59.8	70.7	48.1	69.3	83.8	51.7	58.8
Reasoning FT (200K)	29.9	58.4	68.9	47.7	72.2	84.5	55.9	59.6
reasoning subset	28.8	59.7	66.3	51.4	73.4	90.8	61.4	61.7
reasoning subset 200K	28.3	59.1	66.2	51.1	74.3	90.4	59.9	61.3
instruction subset	27.9	60.8	74.0	48.2	71.1	83.7	51.5	59.6
instruction subset 200K	28.0	59.2	70.7	47.5	73.6	84.1	51.9	59.3
Perception FT	18.0	58.7	69.5	53.7	67.2	97.8	61.5	60.9
Embodiment FT	20.2	63.2	65.8	43.9	62.6	70.9	59.7	55.2
Simulation FT	35.2	60.0	66.6	48.2	72.6	92.0	60.9	62.2
ReasonGraft (3:7)	33.1	59.6	70.0	48.5	72.2	91.4	60.3	62.2
Other open-source models								
LTX-2	23.5	60.1	70.9	42.9	59.8	73.1	55.7	55.2
CogVideoX1.5-5B	39.1	49.5	42.5	60.0	87.1	90.3	95.1	66.2
HunyuanVideo-I2V	25.3	51.8	39.3	50.8	83.4	94.9	76.1	60.2
Wan2.2-TI2V-5B	26.0	47.6	42.4	54.1	88.9	92.9	93.8	63.7

Table 13: World simulation: the Simulation-Calibrated aggregation of Equation 5.

Model / data	A_{cam}	**Ability**	**Consistency**	**Calibrated**
Wan2.2-I2V-A14B				
Baseline	50.0	52.7	83.6	62.0
Reasoning FT (10K)	78.0	60.4	84.2	67.5
Reasoning FT	85.0	61.7	84.8	68.6
high-noise only	77.8	59.9	84.3	67.2
low-noise only	54.6	54.2	84.0	63.2
Reasoning FT (200K)	72.4	58.1	87.1	66.8
reasoning subset	57.5	53.7	89.7	64.5
reasoning subset 200K	66.1	56.6	88.3	66.1
instruction subset	82.8	61.3	85.5	68.5
instruction subset 200K	81.9	60.7	84.9	68.0
Reasoning full FT	43.3	46.4	86.5	58.5
Perception FT	40.4	49.1	90.1	61.4
Embodiment FT	74.4	59.2	85.6	67.1
Simulation FT	98.5	63.7	82.8	69.4
ReasonGraft (3:7)	92.4	63.7	83.7	69.7
ReasonGraft (5:5)	95.7	64.3	83.3	70.0
ReasonGraft (7:3)	97.2	63.9	83.3	69.7
Mixed-data FT (3:7)	98.9	63.4	83.8	69.5
LTX-2.3-I2V				
Baseline	50.0	57.9	60.6	58.7
Reasoning FT	81.6	65.0	68.3	66.0
Reasoning FT (200K)	89.2	66.0	70.8	67.5
reasoning subset	84.1	65.4	75.2	68.3
reasoning subset 200K	81.5	64.5	74.9	67.6
instruction subset	79.3	65.6	68.8	66.5
instruction subset 200K	79.7	64.3	69.9	66.0
Perception FT	6.7	47.1	75.5	55.6
Embodiment FT	14.7	46.9	64.4	52.2
Simulation FT	98.6	68.3	75.2	70.4
ReasonGraft (3:7)	96.8	68.7	74.6	70.5

Table 14: VIPER: POC@1 (correct outcome with a consistent process), outcome correctness (OC), the rate of correct outcomes reached through an inconsistent process (Hack, lower is better), and POC@1 per reasoning domain.

Model / data	**POC@1**	**OC**	**Hack**	**Physics**	**Planning**	**Spatial**	**Structural**	**Symbolic**	**Temporal**
Wan2.2-I2V-A14B									
Baseline	3.0	9.0	6.0	9.1	0.0	0.0	9.1	0.0	0.0
Reasoning FT (10K)	6.0	21.0	15.0	0.0	7.1	0.0	13.6	0.0	14.3
Reasoning FT	8.0	22.0	14.0	9.1	14.3	0.0	9.1	0.0	21.4
high-noise only	5.0	18.0	13.0	9.1	7.1	0.0	4.6	0.0	14.3
low-noise only	1.0	8.0	7.0	0.0	0.0	0.0	4.6	0.0	0.0
Reasoning FT (200K)	5.0	20.0	15.0	0.0	7.1	0.0	9.1	0.0	14.3
Reasoning FT (500K)	10.0	27.0	17.0	9.1	7.1	5.0	9.1	0.0	35.7
Reasoning full FT	5.0	16.0	11.0	9.1	0.0	0.0	4.6	0.0	21.4
Perception FT	7.0	11.0	4.0	18.2	0.0	10.0	13.6	0.0	0.0
Embodiment FT	3.0	13.0	10.0	9.1	0.0	0.0	9.1	0.0	0.0
Simulation FT	2.0	8.0	6.0	9.1	0.0	0.0	4.6	0.0	0.0
LTX-2.3-I2V									
Baseline	1.0	7.0	6.0	0.0	7.1	0.0	0.0	0.0	0.0
Reasoning FT	4.0	15.0	11.0	9.1	7.1	0.0	4.6	0.0	7.1
Perception FT	1.0	4.0	3.0	0.0	7.1	0.0	0.0	0.0	0.0
Embodiment FT	4.0	11.0	7.0	0.0	14.3	5.0	4.6	0.0	0.0
Simulation FT	1.0	10.0	9.0	0.0	7.1	0.0	0.0	0.0	0.0

Table 15: Gen-ViRe: the six reasoning dimensions and their average.

Model / data	Abstract	Alg. & logic	Analogy	Perceptual	Planning	Spatial	Avg.
Wan2.2-I2V-A14B							
Baseline	14.4	33.5	3.3	20.9	59.7	43.0	29.1
Reasoning FT (10K)	40.8	52.5	6.7	36.1	69.1	52.9	43.0
Reasoning FT	45.8	43.0	25.0	37.6	66.9	52.6	45.1
high-noise only	51.9	41.0	25.0	33.7	70.6	49.6	45.3
low-noise only	18.6	37.2	10.0	26.2	62.2	45.9	33.4
Reasoning FT (200K)	54.9	41.7	26.7	34.4	69.8	50.1	46.2
Reasoning FT (500K)	44.4	44.1	30.0	34.2	64.0	49.8	44.4
Reasoning full FT	50.4	50.5	15.0	32.0	59.5	48.7	42.7
Perception FT	21.5	52.0	1.7	35.7	60.8	43.9	35.9
Embodiment FT	14.4	37.9	8.3	26.2	64.2	42.5	32.3
Simulation FT	13.1	31.5	8.3	20.6	57.8	38.4	28.3
LTX-2.3-I2V							
Baseline	15.8	30.9	13.3	18.4	58.8	36.8	29.0
Reasoning FT	45.6	36.6	26.7	33.3	61.3	34.7	39.7
Perception FT	19.2	52.6	16.7	23.3	50.5	45.8	34.7
Embodiment FT	23.3	40.4	21.7	17.1	58.2	37.3	33.0
Simulation FT	14.8	19.1	10.0	6.2	55.0	34.6	23.3

Table 16: V-ReasonBench: pass@1 pooled within each of the four dimensions, and the dimension average.

Model / data	Structured	Spatial	Pattern	Physical	Average
Wan2.2-I2V-A14B					
Baseline	0.0	0.0	0.0	16.7	4.2
Reasoning FT (10K)	0.0	8.8	0.0	6.7	3.9
Reasoning FT	0.6	8.8	5.5	26.7	10.4
high-noise only	1.1	11.8	3.6	30.0	11.6
low-noise only	0.0	1.5	0.0	10.0	2.9
Reasoning FT (200K)	0.0	16.2	10.9	40.0	16.8
Reasoning FT (500K)	0.0	14.7	5.5	46.7	16.7
Reasoning full FT	0.0	22.1	0.0	40.0	15.5
Perception FT	0.0	0.0	0.0	10.0	2.5
Embodiment FT	0.0	1.5	0.0	16.7	4.5
Simulation FT	0.0	1.5	0.0	16.7	4.5
LTX-2.3-I2V					
Baseline	1.7	1.5	0.0	20.0	5.8
Reasoning FT	2.4	7.4	3.6	16.7	7.5
Perception FT	0.0	1.5	0.0	13.3	3.7
Embodiment FT	1.7	0.0	0.0	20.0	5.4
Simulation FT	1.7	1.5	0.0	20.0	5.8

Table 17: VideoThinkBench: last-frame accuracy of the four task families and their mean.

Model / data	ARC-AGI	Eyeballing	Maze	Visual puzzle	Mean
Wan2.2-I2V-A14B					
Baseline	18.9	16.7	0.0	96.0	32.9
Reasoning FT (10K)	56.1	32.9	0.0	100.0	47.2
Reasoning FT	56.1	28.1	0.0	100.0	46.1
high-noise only	55.1	25.2	0.0	100.0	45.1
low-noise only	19.1	25.2	0.0	96.0	35.1
Reasoning FT (200K)	58.9	31.4	0.0	100.0	47.6
Reasoning FT (500K)	54.9	33.8	0.0	100.0	47.2
Reasoning full FT	57.0	27.6	0.0	100.0	46.2
Perception FT	4.1	30.0	0.0	99.0	33.3
Embodiment FT	18.5	21.9	0.0	100.0	35.1
Simulation FT	24.7	14.3	0.0	94.0	33.2
LTX-2.3-I2V					
Baseline	8.4	15.2	0.0	94.0	29.4
Reasoning FT	53.0	33.8	0.0	100.0	46.7
Perception FT	0.6	27.6	0.0	96.0	31.1
Embodiment FT	4.8	20.5	0.0	96.0	30.3
Simulation FT	11.6	21.9	0.0	59.0	23.1

Table 18: Visual-shift scores on VIPER, Gen-ViRe, and VideoThinkBench. VideoThinkBench covers the eyeballing and visual-puzzle families under a VLM judge; All is pooled over their 310 items.

Model / data	VIPER	Gen-ViRe	VTB eyeball.	VTB puzzle	VTB all
Wan2.2-I2V-A14B					
Baseline	5.0	27.4	10.0	19.0	12.9
Reasoning FT	7.0	39.9	7.6	16.0	10.3
high-noise only	8.0	36.2	3.3	18.0	8.1
low-noise only	5.0	29.5	13.8	13.0	13.6
LTX-2.3-I2V					
Baseline	2.0	27.4	11.4	15.0	12.6
Reasoning FT	4.0	35.7	8.6	38.0	18.1

Table 19: Visual-shift V-ReasonBench: pass@1 on the six VLM-assessed tasks and item-weighted pass@1 over their 193 inputs.

Model / data	Math	Sudoku	Code	Shape fit	Color conn.	Temp.	All
Wan2.2-I2V-A14B							
Baseline	0.0	0.0	0.0	0.0	0.0	30.0	1.6
Reasoning FT	0.0	0.0	0.0	21.4	0.0	30.0	4.7
high-noise only	0.0	0.0	0.0	25.0	0.0	40.0	5.7
low-noise only	0.0	0.0	0.0	0.0	0.0	30.0	1.6
LTX-2.3-I2V							
Baseline	0.0	0.0	0.0	0.0	0.0	20.0	1.0
Reasoning FT	0.0	0.0	6.7	10.7	0.0	30.0	4.7

Table 20: Visual-shift VBVR-Bench with the Blender re-rendering. Overall is the mean of the ID and OOD splits; the remaining columns are the ID/OOD × reasoning/instruction cells and the two task-type marginals, as in Table 9.

Model / data	**Overall**	**ID**	**OOD**	**ID**		**OOD**		**All**	
				Reas.	Instr.	Reas.	Instr.	Reas.	Instr.
Wan2.2-I2V-A14B									
Baseline	30.1	33.1	27.1	33.3	32.8	30.1	26.1	32.4	28.3
Reasoning FT	38.0	42.6	33.4	42.8	42.1	38.9	31.5	41.7	35.1
high-noise only	37.5	42.6	32.3	42.6	42.6	38.0	30.3	41.2	34.5
low-noise only	30.1	33.3	27.0	33.5	32.9	30.2	25.9	32.5	28.3
LTX-2.3-I2V									
Baseline	28.7	31.2	26.2	29.9	33.4	30.3	24.8	30.0	27.7
Reasoning FT	33.0	35.7	30.2	35.2	36.7	35.8	28.3	35.4	31.2

Table 21: Visual-shift Gen-ViRe: the six reasoning dimensions and their average.

Model / data	**Abstract**	**Alg. & logic**	**Analogy**	**Perceptual**	**Planning**	**Spatial**	**Avg.**
Wan2.2-I2V-A14B							
Baseline	17.3	30.7	6.7	15.5	54.4	40.1	27.4
Reasoning FT	42.9	33.8	30.0	21.9	57.0	53.9	39.9
high-noise only	40.2	32.6	21.7	21.6	54.8	46.3	36.2
low-noise only	19.8	36.0	6.7	14.1	53.3	47.1	29.5
LTX-2.3-I2V							
Baseline	18.5	33.5	11.7	15.3	46.4	38.9	27.4
Reasoning FT	48.1	40.4	15.0	16.3	51.3	43.2	35.7

Table 22: VBench I2V quality dimensions (Huang et al., 2026b) for the base model and the model after LoRA fine-tuning on 50K clips of XVReason, on both backbones: 355 I2V prompts, one video per prompt, scores scaled to $[0, 100]$.

	Wan2.2-I2V-A14B		**LTX-2.3-I2V**	
Dimension	Baseline	Reasoning FT	Baseline	Reasoning FT
I2V subject	98.3	98.8	97.7	98.0
I2V background	98.6	99.3	98.1	98.4
Subject consistency	95.1	96.3	94.4	94.8
Background consistency	95.5	97.9	97.4	96.9
Aesthetic quality	64.4	66.9	60.7	63.2
Imaging quality	70.7	70.4	70.5	72.1
Temporal flickering	96.8	97.9	98.1	97.9
Motion smoothness	98.1	98.8	98.7	98.4
Dynamic degree	43.3	32.1	29.3	24.8
Mean	84.5	84.3	82.8	82.7