COGOMNICONTROL: REASONING-DRIVEN CONTROLLABLE VIDEO GENERATION VIA CREATIVE INTENT COGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based video generative models achieve strong photorealism and motion fluency, yet fall short in professional production, where conditions are abstract, sparse and often mutually conflicting, e.g., storyboard sketches and clay-render clips that specify the creative intent rather than the pixels. We present **CogOmniControl**, a closed-loop reasoning–generation–verification framework for controllable video generation. The reasoner, CogDirector, is a specialized VLM trained via SFT and RFT on authentic production data; it cognizes the creative intent behind sparse and abstract conditions into a dense production plan, whose representations are injected into generator as semantic guidance through in-context conditioning. The same frozen CogDirector serves as the reward model during generator RFT, anchoring training and test-time verification to a consistent interpretation of creative intent. At test time, it further emits an evaluator harness in the same forward pass to select among multiple sampled candidates via VLM judges and specialized tools, i.e., test-time scaling. We theoretically characterize this verification process, deriving bounds on the misranking probability and expected selection regret of best-of-N selection, and showing that input-adaptive evaluation admits tighter guarantees when it reduces effective evaluation noise while preserving the target quality. We also introduce CogReasonBench and CogControlBench built from real professional workflow data with genuine creative intent. Experiments on the two benchmarks show that CogOmniControl surpasses existing open-source models.

1 INTRODUCTION

Diffusion-based video generative models (Hong et al., 2022; Yang et al., 2024; HaCohen et al., 2024; Wan et al., 2025) have reached striking photorealism and motion fluency. The field is now moving toward omni-level controllable generation (Jiang et al., 2025; Wei et al., 2025; Chen et al., 2026; Pan et al., 2026), which aims at a single system for multimodal inputs and diverse control conditions. Existing systems already handle explicit, pixel-aligned conditions such as pose, depth and masks (Zhang et al., 2023; Jiang et al., 2025). However, professional production remains out of reach: the conditions there are abstract, sparse and often mutually conflicting, e.g., storyboard sketches and clay-render clips that specify the creative intent rather than the pixels (Fig. 1). There is a research question: *how can a video generator acquire the professional creative cognition that intent-level conditions demand?*

Two gaps stand in the way. ① **Cognitive Gap**: confronted with complex or even conflicting control signals in professional workflows, current VLMs fail to comprehend the underlying creative intent and to formulate generation plans grounded in domain-specific creative knowledge. ② **Alignment Gap**: it remains open whether the reasoning produced under abstract conditions is properly aligned with the generated video, and the reasoning of a generic VLM further injects noise into generation (Yang et al., 2026; Chen et al., 2026).

We present **CogOmniControl**, a closed-loop reasoning–generation–verification framework. Given heterogeneous conditions, **CogDirector** first cognizes their underlying creative intent and converts sparse and abstract cues into a dense production reasoning. Generator then combines this intent-level reasoning with heterogeneous pixel-level conditions through in-context conditioning to generate

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.

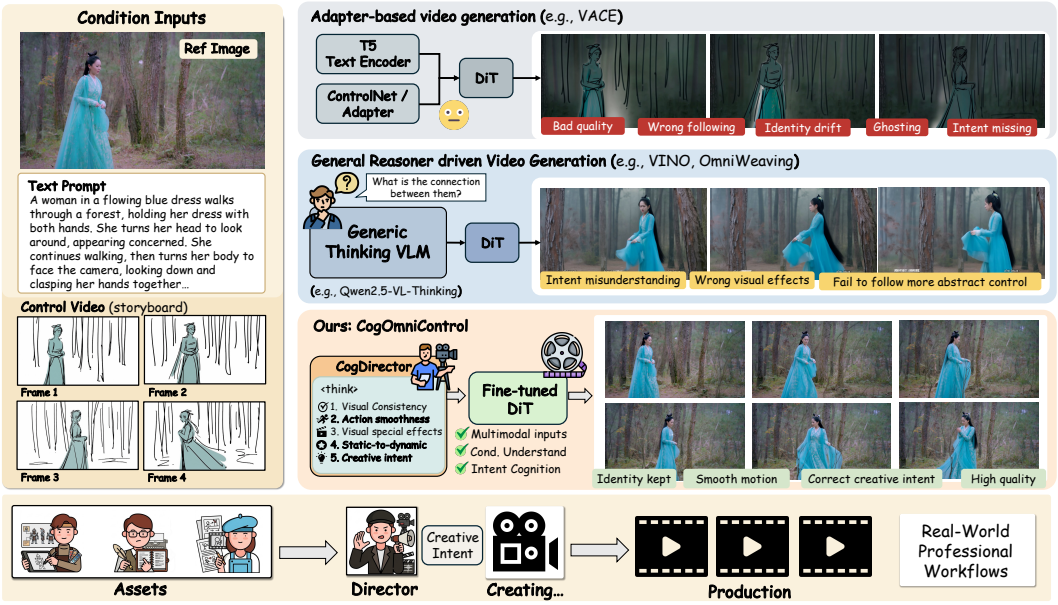

Figure 1: The motivation of our CogOmniControl. The adapter-based methods and video generation models with generic VLMs fail to generate the final video from the given condition.

the video. At inference, CogDirector additionally produces an input-adaptive evaluator harness for test-time candidate selection.

CogDirector is obtained from a generic VLM through SFT followed by RFT on authentic professional production data. Its reasoning representations provide semantic guidance to generator, while the same frozen CogDirector serves as the reward model during generator RFT. This stage-wise design keeps the reasoning distribution, semantic conditioning, and reward criterion fixed while optimizing the generator, thereby avoiding a moving conditioning and evaluation target. Specifically, CogDirector predicts the production reasoning $\mathcal{R}$ and evaluator harness $\mathcal{H}$ in a single forward pass. Generator samples N candidates under $(\mathcal{R}, \mathcal{C})$, and the harness selects $\mathcal{V}^* = \arg\max_{\mathcal{V}_i} S(\mathcal{V}_i; \mathcal{H})$. Rather than using a fixed evaluator suite for all inputs, CogDirector selects and weights evaluators according to the inferred creative intent; for example, identity consistency is unnecessary when no identity reference is provided.

Beyond the architectural design, we provide a theoretical characterization of the closed-loop verification process. We model the harness score as a noisy estimate of an underlying target quality and show that the effective evaluator noise controls the reliability of best-of-N selection. Specifically, we derive upper bounds on both the probability of misranking the truly best candidate and the corresponding expected selection regret. The analysis further shows that, when an input-adaptive harness preserves the same target quality while reducing effective evaluation noise, it admits strictly tighter bounds than a fixed evaluator suite.

Evaluating such a pipeline further requires benchmarks that carry genuine creative intent, which existing ones lack because they are built from simulated intent. We therefore introduce two benchmarks, CogReasonBench and CogControlBench, which evaluate the intent cognition of CogDirector and the intent-aligned generation of CogOmniControl, respectively; both are built from storyboards, clay-render clips and their final videos collected in real-world professional production. Experiments on the two benchmarks show that CogOmniControl outperforms existing open-source models.

Our contribution can be summarized as follows:

- We formulate intent-level controllable video generation for professional production, where conditions can be sparse, abstract, and mutually conflicting, and present **CogOmniControl**, a closed-loop reasoning–generation–verification framework that explicitly decomposes creative-intent cognition from conditional video generation.
- We develop **CogDirector**, a specialized VLM trained with SFT and RFT on authentic production data, together with the video generator, which injects CogDirector's intent-level reasoning representations alongside heterogeneous visual conditions. During generator RFT, the same frozen

CogDirector provides the reward criterion, yielding a stage-wise optimization process with stationary reasoning, conditioning, and evaluation criteria.

- We introduce an **input-adaptive evaluator harness** for test-time scaling. CogDirector jointly predicts production reasoning and the evaluator harness, which dynamically selects and weights VLM judges and specialized evaluators according to each input and performs best-of-N selection.
- We theoretically characterize the proposed verification mechanism, deriving bounds on best-of-N misranking probability and expected selection regret under evaluator noise, and identifying the condition under which input-adaptive evaluation has tighter guarantees than a fixed suite. We further construct **CogReasonBench** and **CogControlBench** from authentic professional workflows to evaluate intent cognition and intent-aligned controllable generation.

2 Related Work

Diffusion models produce high-fidelity image and video content (Rombach et al., 2022; Peebles & Xie, 2023; Hong et al., 2022; Yang et al., 2024; Kong et al., 2024; Wan et al., 2025). For precise control, adapters inject pixel-aligned conditions (Zhang et al., 2023; Li et al., 2025b; Jiang et al., 2025; Liu et al., 2025a), while unified omni-level generation is pursued in the image domain (Xiao et al., 2025; Wu et al., 2025a) and extended to video by proprietary systems (Seedance et al., 2026; Team et al., 2025; OpenAI, 2025; Google, 2025) and open-source efforts (Jiang et al., 2025; Wei et al., 2025; Chen et al., 2026; Pan et al., 2026), which, however, lack a deep understanding of diverse conditions and creative intent. RL for visual generation advances from preference optimization (Rafailov et al., 2023; Black et al., 2024) to GRPO-based training of flow-matching models (Shao et al., 2024; Liu et al., 2025b; Xue et al., 2025). Meanwhile, harness engineering (Gao et al., 2024; Lin et al., 2026) and VLM-as-a-Judge (Zheng et al., 2023; Chen et al., 2024) scale the capability of a fixed model at inference time for candidate selection; yet existing protocols apply a fixed evaluator set to all samples, whereas CogDirector adaptively assigns evaluators per input. In contrast to prior harness-based selection schemes that are primarily empirical, we additionally analyze candidate selection under evaluator noise and characterize the conditions under which adaptive evaluation provides tighter selection guarantees. The complete related work is provided in Appendix B.

3 Method

3.1 CogOmniControl Framework

Given multimodal conditions $\mathcal{C} = (V_{\text{ctrl}}, I_{\text{ref}}, T_{\text{desc}})$, CogOmniControl first employs CogDirector to infer an intermediate production reasoning $\mathcal{R}$ and then conditions generator on both $\mathcal{C}$ and $\mathcal{R}$ to generate the final video. Here, V_{ctrl}, I_{ref}, and T_{desc} provide spatial-temporal, visual-reference, and semantic information, respectively. We introduce $\mathcal{R}$ as an intermediate reasoning variable, which decomposes controllable generation into cognition and generation:

$$p(\mathcal{V} \mid \mathcal{C}) = \mathbb{E}_{\mathcal{R}\sim\pi_\phi(\cdot|\mathcal{C})}\left[p_\theta(\mathcal{V} \mid \mathcal{R}, \mathcal{C})\right], \tag{1}$$

This latent-variable factorization explicitly separates creative-intent cognition from conditional generation: π_ϕ infers the production reasoning, whereas p_θ renders a video conditioned on both the reasoning and the original conditions. At the population level, an ideal objective is to maximize

$$J(\phi, \theta) = \mathbb{E}_{\mathcal{C}\sim\mathcal{D},\, \mathcal{R}\sim\pi_\phi(\cdot|\mathcal{C}),\mathcal{V}\sim p_\theta(\cdot|\mathcal{R},\mathcal{C})}\left[Q(\mathcal{V}; \mathcal{R}, \mathcal{C})\right], \tag{2}$$

where $\mathcal{D}$ denotes the data distribution. Here, Q denotes an ideal population-level task utility used only to formalize the target problem; it is not directly available during training. Direct optimization would require repeatedly evaluating downstream video quality for sampled reasoning trajectories and is computationally prohibitive. We therefore optimize two task-aligned surrogate objectives stage by stage:

$$\begin{aligned} J_{\text{R}}(\phi) &= \mathbb{E}_{\mathcal{C}\sim\mathcal{D},\, \mathcal{R}\sim\pi_\phi(\cdot|\mathcal{C})}\left[R_{\text{CogDirector}}(\mathcal{R}, \mathcal{C})\right], \\ \phi^* &= \arg\max_\phi J_{\text{R}}(\phi), \\ J_{\text{G}}(\theta; \phi^*) &= \mathbb{E}_{\mathcal{C}\sim\mathcal{D},\, \mathcal{R}\sim\pi_{\phi^*}(\cdot|\mathcal{C}),\mathcal{V}\sim p_\theta(\cdot|\mathcal{R},\mathcal{C})}\left[R_{\text{visual}}(\mathcal{V}, \mathcal{R}, \mathcal{C})\right]. \end{aligned} \tag{3}$$

The two stages thus optimize task-aligned surrogate objectives for reasoning quality and intent-aligned generation, respectively. We do not assume that these surrogates are equivalent to the ideal objective in Eq. 2.

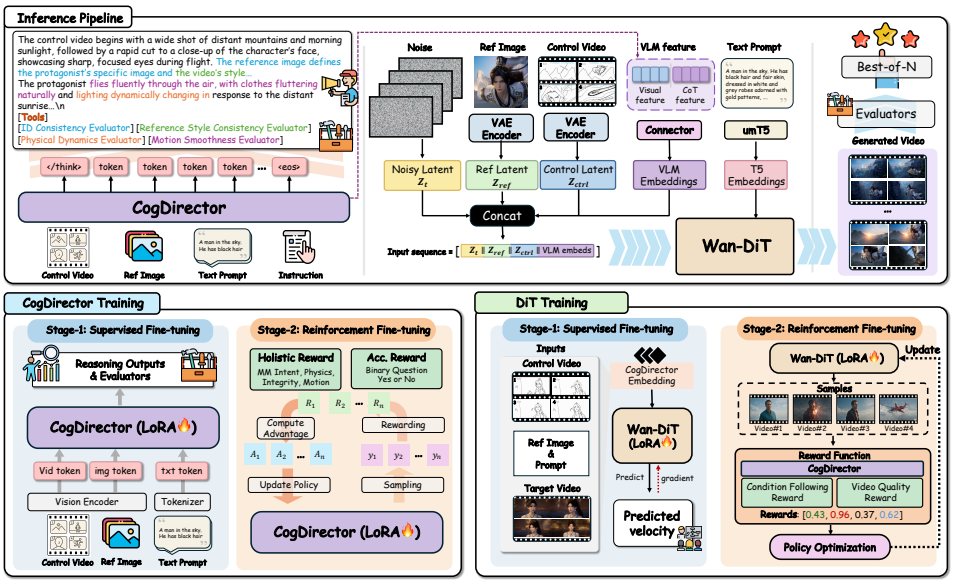

Figure 2: Overview of CogOmniControl. CogDirector interprets the multimodal conditions and produces both production reasoning and an input-adaptive evaluator harness. The reasoning representation is injected into the generator together with the visual conditions. During training, the frozen CogDirector further provides the reward criterion for generator RFT. At inference, multiple candidates are sampled and ranked by the input-adaptive harness for test-time scaling.

Lemma 1 (Stationarity of the CogDirector-dependent criterion)**.** *During generator RFT, let CogDirector be fixed at ϕ^*. Then the reasoning distribution $\pi_{\phi^*}(\mathcal{R}|\mathcal{C})$, the semantic-conditioning mapping produced by CogDirector, and the parameters defining the CogDirector-based reward criterion are invariant to updates of the generator parameters θ. Consequently, generator optimization changes the generated-video distribution while the CogDirector-dependent conditioning and evaluation criteria remain fixed.*

Note that the realized reward value may still change with the generated video; the lemma concerns stationarity of the reward criterion rather than constancy of its outputs. This stationarity does not by itself imply convergence or monotonic improvement of finite-sample RFT. The proof is provided in Appendix A.1. At inference, CogDirector jointly produces the reasoning $\mathcal{R}$ and an evaluator harness $\mathcal{H}$. The resulting best-of-N verification procedure is described in Sec. 3.4.

3.2 CogDirector: Cognizing Creative Intent from Multimodal Conditions

CogDirector maps heterogeneous conditions $\mathcal{C}$ into a dense production reasoning $\mathcal{R}$. It identifies cross-modal correspondences, resolves potential conflicts, and infers production-relevant information not explicitly specified by any individual condition, including event ordering, scene layout, motion, lighting, and physical effects. For example, combining "raining" in the textual description with standing water in a reference image may imply visible ripple effects. The resulting reasoning is subsequently used both as intent-level guidance for the generator and as the semantic basis for training-time and test-time evaluation.

Training. We train CogDirector using SFT and RFT. SFT on authentic production data endows CogDirector with the output format of production reasoning and the domain-specific knowledge of professional workflows, and the subsequent RFT further refines the reasoning quality with a verifiable reward. For RFT, we combine a holistic reasoning reward with an atomic fact-verification reward:

$$R_{\text{holistic}} = \sum_{k \in \mathcal{K}} w_k \, \text{VLM}_k(\mathcal{R}, \mathcal{C}), \;\; R_{\text{acc}} = \frac{1}{M} \sum_{j=1}^{M} \text{VLM}(\mathcal{R}, q_j), \;\; R_{\text{CogDirector}} = R_{\text{holistic}} + \lambda R_{\text{acc}}, \quad (4)$$

where $\mathcal{K} = \{\text{intent}, \text{phys}, \text{info}, \text{dyn}\}$ corresponds to creative intent, physical plausibility, information integrity, and motion description. The holistic term $\text{VLM}_k(\cdot)$ denotes the normalized score assigned by the judge model for dimension k under the LLM-as-a-Judge protocol (Chen et al., 2024). The atomic questions $\{q_j\}_{j=1}^{M}$ are generated from the input conditions to verify factual consistency of the reasoning, which transforms subjective narrative evaluation into a verifiable accuracy metric.

We optimize Eq. 4 using GRPO (Shao et al., 2024) with group-relative advantages. Detailed optimization settings and the clipped sample-level objective are provided in the Appendix.

3.3 Generator: Empowering Video Diffusion Transformer via CogDirector

To transfer CogDirector's intent-level reasoning into the generator, we project its reasoning representation through a lightweight connector and concatenate it with the visual latent sequence:

$$X_t = \mathrm{Concat}\left(Z_t, Z_{\mathrm{ref}}, Z_{\mathrm{ctrl}}, E_{\mathcal{R}}\right), \tag{5}$$

where Z_t is the noisy video latent, Z_{ref} and Z_{ctrl} are the reference-image and control-video latents, respectively, and $E_{\mathcal{R}}$ is the CogDirector reasoning representation after the connector. The textual description T_{desc} is incorporated into CogDirector's reasoning and therefore enters DiT through $E_{\mathcal{R}}$.

We additionally use the same trained CogDirector as the reward model for generator RFT:

$$R_{\mathrm{visual}} = \sum_{m \in \mathcal{M}} w_m \mathrm{VLM}_m(\mathcal{V}, \mathcal{R}, \mathcal{C}), \tag{6}$$

where $\mathcal{M} = \{\text{condition following}, \text{video quality}\}$. The RFT is performed at lower resolution, and inference is performed at high resolution due to the scaling capability of video diffusion transformer (Ping et al., 2025; 2026). During generator RFT, CogDirector remains frozen; only the DiT LoRA parameters and the lightweight connector are optimized. The resulting stationary CogDirector-dependent conditioning and reward criterion follows Lemma 1.

3.4 Closed-Loop Verification with Evaluator Harness

A fixed evaluator suite applies the same criteria to every generation regardless of the input intent. However, different condition sets require different evaluation dimensions; for example, identity consistency is irrelevant when no identity reference is provided. We therefore let CogDirector jointly predict the production reasoning $\mathcal{R}$ and an input-adaptive evaluator harness $\mathcal{H}$ in a single forward pass:

$$(\mathcal{R}, \mathcal{H}) \sim \pi_{\phi^*}(\cdot \mid \mathcal{C}), \qquad \mathcal{H} = \{(e_k, w_k)\}_{k=1}^{K}, \tag{7}$$

where e_k is the k-th evaluator selected from the `[tools]` library and w_k is its weight, with each evaluator returning a normalized score in $[0, 1]$. Please refer to the Appendix for the details of these designed tools. Given $(\mathcal{R}, \mathcal{C})$, Generator samples N candidates $\{\mathcal{V}_i\}_{i=1}^{N}$ from different initial noise realizations. The harness assigns

$$S_i = S(\mathcal{V}_i; \mathcal{H}) = \sum_{k=1}^{K} w_k e_k(\mathcal{V}_i, \mathcal{R}, \mathcal{C}), \qquad \sum_{k=1}^{K} w_k = 1, \tag{8}$$

and selects

$$\mathcal{V}^* = \arg\max_{1 \le i \le N} S_i. \tag{9}$$

Assumption 1 (Noisy evaluator model)**.** *For theoretical analysis, let q_{ik} denote the latent ground-truth quality of candidate i under evaluation criterion k, and let*

$$e_k(\mathcal{V}_i, \mathcal{R}, \mathcal{C}) = q_{ik} + \xi_{ik},$$

where ξ_{ik} is the corresponding evaluation error. For a fixed harness $\mathcal{H}$, define its latent target quality as

$$Q_i^{\mathcal{H}} = \sum_{k=1}^{K} w_k q_{ik}.$$

Since $\sum_{k=1}^{K} w_k = 1$, the observed harness score satisfies

$$S_i = Q_i^{\mathcal{H}} + \varepsilon_i, \qquad \varepsilon_i = \mathbf{w}^\top \boldsymbol{\xi}_i, \tag{10}$$

where $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^\top$. We assume $\boldsymbol{\xi}_i$ is zero-mean jointly sub-Gaussian with proxy matrix $\Sigma_{\mathcal{H}}$, i.e.,

$$\mathbb{E}\left[\exp\left(\mathbf{u}^\top \boldsymbol{\xi}_i\right)\right] \le \exp\left(\frac{1}{2}\mathbf{u}^\top \Sigma_{\mathcal{H}} \mathbf{u}\right), \qquad \forall \mathbf{u}. \tag{11}$$

Lemma 2 (Noise concentration of a weighted evaluator harness)**.** *Under Assumption 1, the aggregated harness error $\varepsilon_i = \mathbf{w}^\top \boldsymbol{\xi}_i$ is sub-Gaussian with variance proxy*

$$\sigma_{\mathcal{H}}^2 = \mathbf{w}^\top \Sigma_{\mathcal{H}} \mathbf{w}. \tag{12}$$

If evaluator errors are mutually uncorrelated with individual proxies σ_k^2, this reduces to $\sigma_{\mathcal{H}}^2 = \sum_{k=1}^K w_k^2 \sigma_k^2$.

The proof is provided in Appendix A.2.

Proposition 1 (Reliability of best-of-N harness selection)**.** *Let $i^* = \arg\max_i Q_i^{\mathcal{H}}$ be the truly best candidate and define*

$$\Delta = Q_{i^*}^{\mathcal{H}} - \max_{j \neq i^*} Q_j^{\mathcal{H}} > 0, \qquad B = Q_{i^*}^{\mathcal{H}} - \min_j Q_j^{\mathcal{H}}. \tag{13}$$

Assume that, for every $j \neq i^*$, the score-error difference $\varepsilon_j - \varepsilon_{i^*}$ is zero-mean sub-Gaussian with variance proxy $\nu_{\mathcal{H}}^2$. Then

$$\mathbb{P}[\text{misrank}] \leq (N-1) \exp\left(-\frac{\Delta^2}{2\nu_{\mathcal{H}}^2}\right). \tag{14}$$

Moreover,

$$\mathbb{E}\left[Q_{i^*}^{\mathcal{H}} - Q^{\mathcal{H}}(\mathcal{V}^*)\right] \leq B(N-1) \exp\left(-\frac{\Delta^2}{2\nu_{\mathcal{H}}^2}\right). \tag{15}$$

If the candidate-level errors are independent and each is sub-Gaussian with proxy $\sigma_{\mathcal{H}}^2$ from Lemma 2, then $\nu_{\mathcal{H}}^2 \leq 2\sigma_{\mathcal{H}}^2$ and therefore

$$\mathbb{P}[\text{misrank}] \leq (N-1) \exp\left(-\frac{\Delta^2}{4\sigma_{\mathcal{H}}^2}\right), \tag{16}$$

with the expected regret bounded accordingly by

$$\mathbb{E}\left[Q_{i^*}^{\mathcal{H}} - Q^{\mathcal{H}}(\mathcal{V}^*)\right] \leq B(N-1) \exp\left(-\frac{\Delta^2}{4\sigma_{\mathcal{H}}^2}\right). \tag{17}$$

The proof is provided in Appendix A.2.

Proposition 1 reveals a simple signal-to-noise principle: best-of-N selection is reliable when the quality gap between the best and competing candidates is sufficiently large relative to evaluator noise. Increasing N enlarges the candidate pool but simultaneously introduces more opportunities for noisy misranking, so test-time scaling requires both candidate diversity and sufficiently reliable evaluation.

Corollary 1 (Advantage of lower-noise adaptive evaluation)**.** *Consider an adaptive harness and a fixed evaluator suite applied to the same candidate set. Suppose both estimate the same latent target quality $Q_i^{\mathcal{H}}$ without systematic bias, but their score-error differences have proxies ν_{adaptive}^2 and ν_{fixed}^2, respectively. If $\nu_{\text{adaptive}}^2 < \nu_{\text{fixed}}^2$, then the adaptive harness has strictly tighter upper bounds on both misranking probability and expected selection regret.*

Thus, adaptivity is theoretically beneficial *when* removing intent-irrelevant or noisy evaluators decreases effective score noise without changing the target utility being estimated.

4 BENCHMARK

To further demonstrate the capabilities of CogOmniControl, we curated a new video reasoning and generation benchmark consisting of the storyboard/clay render video and the final videos collected from in-house professional anime production pipelines. This type of data reflects the inherent gap between the abstract condition provided by the user and the raw creative intent in professional production. Additionally, to showcase the generalizability of CogOmniControl in controllable video generation tasks, we incorporated a variety of general controllable generation data, including samples from the community and VACE-Bench (Jiang et al., 2025). Leveraging these data, we built the CogReasonBench to measure the VLM's ability to cognize creative intent and reasoning, and the CogControlBench to measure the quality and condition following of controllable video generation of the model under abstract and sparse conditions.

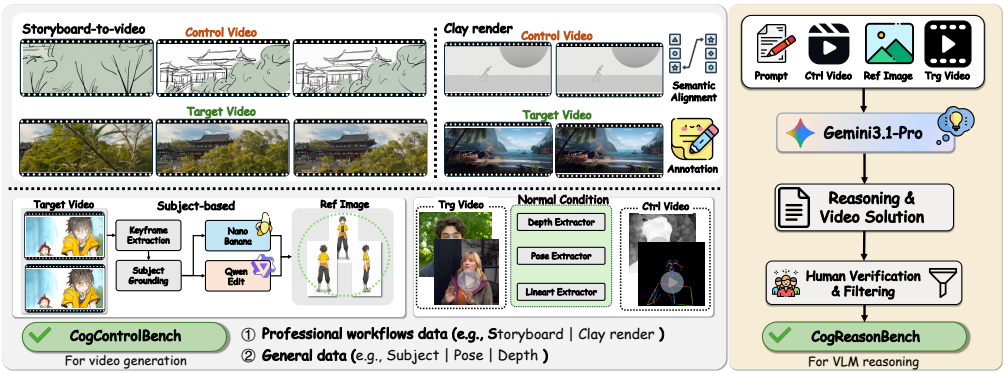

Figure 3: The construction pipeline of CogControlBench and CogReasonBench. We include the professional workflow data and general video generation data in the training set and benchmark.

Table 1: Comparison of CogControlBench with the existing video benchmark.

Benchmark	#Size	Resolution	Input Modality			Data Attributes		VLM-as-a-Judge
			Image	Multi-Image	Video	Condition Source	Creative Intent	
VBench (Huang et al., 2024)	946	480P	✗	✗	✗	-	-	✗
VBench++ (Huang et al., 2025)	2384	480P	✓	✗	✗	-	-	✗
OpenVE-Bench (He et al., 2025a)	431	720P	✗	✗	✓	-	-	✓
TGVE+ (Singer et al., 2024)	1417	480P	✗	✗	✓	Synthetic	Simulated	✗
OpenS2V-Eval (Yuan et al., 2025)	180	720P	✓	✓	✗	Synthetic	Simulated	✗
VACE-Bench (Jiang et al., 2025)	480	480P	✓	✓	✓	Synthetic	Simulated	✗
IntelligentVBench (Pan et al., 2026)	1030	480P	✓	✓	✓	Synthetic	Simulated	✓
CogControlBench	200	720P	✓	✓	✓	**Native**	**Raw**	✓

As shown in Fig. 3, for professional workflow data in CogControlBench, we perform manual semantic alignment and annotation to ensure that the control clip and the final clip share the same semantics. For general data, we incorporate reference-to-video by extracting subjects from key frames and then editing them using Nano-Banana or Qwen-Image-Edit. We also apply condition extractors to extract conditions frame-by-frame to make the dataset support general controllable generation. Tab. 1 shows the comparison of CogControlBench with other video generation benchmarks. To align with the high-quality standards of anime production while optimizing for validation efficiency, we curated a set of 200 high-resolution representative samples. This scale is aligned with established high-resolution benchmarks. For CogReasonBench in VLM, we prompt Gemini3.1-Pro (Google) to reason across the input conditions and the target video to formulate the generative solution. To ensure the correctness of the chain of thought and the solution, the whole process is under human verification and filtering. These two benchmarks and code will be available.

5 Experiment

5.1 Experiment Setup

Experiments are conducted with Qwen3-VL-8B-Thinking (Bai et al., 2025) as the base VLM and Wan2.2-T2V-14B (Wan et al., 2025) as the base DiT on 32 NVIDIA H20 96GB GPUs, where both CogDirector and DiT are trained with LoRA (Hu et al., 2022) via SFT followed by RFT. The training corpus is constructed from the same in-house professional animation production pipeline that provides the benchmark, together with general controllable generation data from the community and public sources, and all clips appearing in CogControlBench and CogReasonBench are excluded from the training set to avoid any leakage. In total, we collect 20k condition–solution pairs for the SFT of CogDirector, 1.2k condition sets for its RFT, and 300k condition–video pairs for the training of DiT, in which the professional workflow data accounts for 32%. The complete training data statistics and implementation details are provided in Appendix C.1. For evaluation, we adopt specialist metrics from VBench (Huang et al., 2024) and a VLM-as-a-Judge (Zheng et al., 2023) protocol with Gemini 3.1-Pro (Google) as the judge, focusing on two dimensions: *condition following*, which measures whether the generated video faithfully realizes the creative intent implied by the condition set $\mathcal{C}$ as a holistic objective, and *visual quality*, which covers aesthetic and imaging quality, temporal flickering, motion smoothness, dynamic degree, identity consistency, and dynamic plausibility. The complete metric definitions are provided in Appendix C.2.

Table 2: The results of CogDirector on CogReasonBench.

Models	MM Intent	Physics	Integrity	Motion	Avg
InternVL-3.5-8B	2.531	3.904	3.847	4.615	3.724
Qwen3-VL-8B-Instruct	2.480	4.045	3.905	4.420	3.712
Qwen3-VL-8B-Thinking	2.670	3.824	3.829	4.727	3.752
CogDirector (SFT)	4.302	4.658	4.623	4.814	4.599
CogDirector (RFT)	**4.572**	**4.766**	**4.810**	**4.942**	**4.772**

Table 3: The comparison on CogControlBench. $\mathcal{AQ}$=Aesthetic Quality, $\mathcal{IQ}$=Image Quality, $\mathcal{TF}$=Temporal Flickering, $\mathcal{MS}$=Motion Smoothness, $\mathcal{DD}$=Dynamic Degree, $\mathcal{MI}$=Multimodal Intent, $\mathcal{AF}$=Appearance Follow, $\mathcal{SF}$=Style Follow, $\mathcal{CF}$=Content Follow, $\mathcal{DF}$=Dynamic Follow, $\mathcal{MN}$=Motion Naturalness, $\mathcal{IC}$=Identity Consistency, $\mathcal{DP}$=Dynamic Plausibility. The **Avg** score is calculated based on 0-1 normalization.

Models	Specialist Metrics					Condition Following					Visual Quality					Avg
	$\mathcal{AQ}$	$\mathcal{IQ}$	$\mathcal{TF}$	$\mathcal{MS}$	$\mathcal{DD}$	$\mathcal{MI}$	$\mathcal{AF}$	$\mathcal{SF}$	$\mathcal{CF}$	$\mathcal{DF}$	$\mathcal{AQ}$	$\mathcal{IQ}$	$\mathcal{MN}$	$\mathcal{IC}$	$\mathcal{DP}$	
Proprietary Models																
Kling-3$\mathcal{O}$ (Team et al., 2025)	0.571	0.644	0.979	0.987	0.511	3.510	4.205	4.267	2.679	**3.526**	3.936	3.453	2.465	3.140	3.203	0.704
Seedance2.0 (Seedance et al., 2026)	0.589	**0.653**	0.980	0.989	0.517	**4.110**	**4.252**	**4.348**	**4.412**	3.054	**4.050**	**3.731**	2.731	3.469	3.494	**0.750**
Open-Source Models																
VACE-Wan2.1 (Jiang et al., 2025)	0.549	0.636	0.975	0.986	0.528	3.421	3.361	3.712	3.886	2.614	3.777	3.680	2.757	3.592	3.330	0.665
VACE-LTX (Jiang et al., 2025)	0.496	0.617	0.980	0.989	0.345	2.807	2.051	1.849	3.377	2.412	2.797	2.588	1.887	2.492	2.299	0.556
VACE-Wan2.2	0.551	0.639	0.980	0.989	0.532	3.518	3.352	3.734	3.900	2.743	3.784	3.687	2.753	3.613	3.114	0.701
VINO (Chen et al., 2026)	0.570	0.581	0.980	0.989	0.280	3.324	3.853	4.020	4.116	2.327	3.855	3.626	2.710	3.341	3.344	0.686
OmniWeaving (Pan et al., 2026)	0.512	0.549	0.976	0.982	0.396	2.630	2.119	2.550	3.963	2.574	3.257	2.941	2.408	3.033	3.000	0.607
CogOmniControl	0.596	0.612	0.980	**0.990**	0.531	3.656	3.786	4.288	4.237	2.684	3.910	3.599	2.856	3.616	3.603	0.730
CogOmniControl (TTS)	0.596	0.632	**0.981**	**0.990**	0.522	3.737	3.890	4.039	4.292	2.806	3.976	3.618	2.741	3.556	3.592	0.731
CogOmniControl (TTS + Harness)	**0.597**	0.641	0.980	**0.990**	**0.536**	3.922	4.028	4.260	4.375	3.013	4.022	3.657	**2.859**	**3.647**	**3.608**	0.748

5.2 Results

Quantitative Results. The results of CogDirector via SFT and RFT are shown in Tab 2, generic VLMs (e.g., Qwen3-VL-8B-Instruct and Thinking) fail to cognize creative intent from multimodal inputs, they also underperform compared to CogDirector in terms of information integrity and motion description. Furthermore, they struggle to generate accurate descriptions for inferred physical effects. Tab. 3 reports the results on CogControlBench. CogOmniControl achieves the highest average score (0.730) among all open-source competitors, surpassing VACE-Wan2.2 (0.701), VINO (0.686) and VACE-Wan2.1 (0.665), while narrowing the gap to the strongest proprietary system Seedance2.0 (0.750) in this task. Further gains are obtained by scaling the test-time compute of the fixed generator: with $N = 4$ sampled candidates, CogOmniControl reaches 0.731 when the candidates are scored by the full set of evaluators, and 0.748 when they are scored by the specific evaluators that CogDirector selects for each individual input. The input-adaptive evaluator selection is consistently better than the fixed one, indicating that CogDirector can effectively serve as the harness of the entire framework. The independent human studies are provided in Appendix B.1.

Qualitative Results. The visual results are presented in Fig. 4 and Fig. 5, the adapter-based methods (e.g., VACE-LTX and VACE) tend to align with control videos at the pixel level, resulting in significant artifacts. Besides, these methods cause semantic misalignment, since clay render videos consist of sparse and abstract control. For the reference-to-video task, CogOmniControl remains strong in performance. The generated videos from VACE lack quality and reference following, while VINO produces virtually static outputs that lack meaningful temporal dynamics.

5.3 Ablation Studies

The ablation studies primarily focus on two questions: (i) how the VLM that provides the multimodal embedding affects the generation of DiT generator, and (ii) how the VLM that serves as the reward model affects the RFT of DiT generator. As shown in Tab. 4, in the DiT-SFT stage, replacing the vanilla Qwen3-VL-8B-Thinking with CogDirector(SFT) and CogDirector(RFT) as the embedding VLM steadily improves the multimodal intent score $\mathcal{MI}$ from 3.142 to 3.401 and 3.582. This demonstrates that a VLM specialized for controllable generation is necessary, since it understands the various conditions and reasons behind the underlying creative intent. In the DiT-RFT stage, using CogDirector(RFT) itself as the reward model further improves $\mathcal{MI}$ from 3.599 to 3.656 and $\mathcal{DP}$ from 3.529 to 3.603, compared with using the generic Qwen3-VL-8B-Thinking as the reward model, confirming that the intent-aware feedback from CogDirector aligns the generator with the creative intent better than a generic judge. The effect of the evaluator harness, which yields 0.731 with the full evaluator set and 0.748 with the adaptively selected one, is reported in Tab. 3. The advantage of input-adaptive evaluation over the full fixed evaluator set is consistent with our theoretical analysis: removing input-irrelevant evaluation dimensions can improve selection when it reduces effective evaluation noise while preserving the relevant quality signal.



Figure 4: The comparison of CogOmniControl with other video generation models in clay render, which is a common intermediate draft stage in workflow production. Zoom in for more details.

Table 4: Ablation studies of CogOmniControl on CogControlBench. We first study the effect of different VLMs as multimodal embeddings for the DiT, and then investigate different VLMs as reward models for DiT RFT.

Stage	Embedding VLM	Reward Model	VLM-as-a-Judge					VLM-as-a-Judge				
			MI	*AF*	*SF*	*CF*	*DF*	*AQ*	*IQ*	*MN*	*IC*	*DP*
DiT-SFT	Qwen3-VL-8B-Thinking	–	3.142	3.541	3.916	4.208	2.467	3.822	3.562	2.708	**3.616**	3.473
	CogDirector(SFT)	–	3.401	3.717	3.990	**4.279**	2.631	3.901	3.553	2.769	3.592	3.490
	CogDirector(RFT)	–	3.582	3.751	4.207	4.172	2.672	3.895	3.594	2.763	3.590	3.514
DiT-RFT	CogDirector(RFT)	Qwen3-VL-8B-Thinking	3.599	3.770	4.268	4.220	**2.689**	3.906	**3.603**	2.849	**3.616**	3.529
	CogDirector(RFT)	**CogDirector(RFT)**	**3.656**	**3.786**	**4.288**	4.237	2.684	**3.910**	3.599	**2.856**	**3.616**	**3.603**

This empirical improvement complements the theoretical characterization in Sec. 3.4, which attributes reliable candidate selection to the quality-gap-to-evaluator-noise ratio.



Figure 5: The comparison of CogOmniControl with other video generation models. Zoom in for more details.

6 Conclusion

In this work, we presented **CogOmniControl**, a reasoning-driven framework for controllable video generation under sparse, abstract, and potentially conflicting professional conditions. CogOmniControl explicitly factorizes the problem into creative-intent cognition and conditional generation: CogDirector converts heterogeneous conditions into dense production reasoning, while the generator integrates the resulting intent-level representation with pixel-level controls and is further aligned through RFT using the same frozen CogDirector. Beyond generation guidance, CogDirector produces an input-adaptive evaluator harness for best-of-N test-time selection. We further provide a theoretical characterization of this verification process by analyzing evaluator noise, deriving bounds on misranking probability and expected selection regret, and identifying the condition under which adaptive evaluation provides tighter guarantees than a fixed evaluator suite. Finally, we introduce CogReasonBench and CogControlBench using authentic professional production data. Experiments demonstrate strong improvements over existing open-source controllable video generators and show that the adaptive harness further benefits test-time scaling.

AI USE STATEMENT

In this work, we used generative AI tools for writing polishing and assistance with code debugging. We have not used generative AI tools for research ideation, or experimental setup, and primary manuscript drafting and data analysis are not applicable to this work as AI-assisted tasks. Additionally, we used generative AI tools for language refinement and literature discovery. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work strictly adheres to the ICLR Code of Ethics. For the human evaluation conducted on our proposed benchmark, all participants were recruited with informed consent and fairly compensated at hourly rates above the local minimum wage. No personally identifiable information or sensitive personal data was collected, and all human responses were anonymized. Regarding dataset construction, all assets included in our benchmark are sourced under appropriate public licenses or fair use guidelines, meticulously filtered to exclude harmful, toxic, or offensive content. The authors declare no conflicts of interest or violations of research integrity.

REPRODUCIBILITY STATEMENT

The complete construction methodology, curation guidelines, and data statistics of our newly introduced benchmark are detailed in the paper. All evaluation pipelines, inference code, and exact computational environment configurations will be released after accepted.

REFERENCES

Vidu AI. Vidu. https://www.vidu.com/, 2026.

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

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YCWjhGrJFD.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. Vino: A unified visual generator with interleaved omnimodal context. *arXiv preprint arXiv:2601.02358*, 2026.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Neural Information Processing Systems*, 36:79858–79885, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Google. Gemini-3. URL https://gemini.google.com. Accessed: 15 Dec. 2025.

Google. Gemini ai video generator powered by veo 3.1. https://gemini.google/overview/video-generation/, 2025.

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

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

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.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. *arXiv preprint arXiv:2512.07826*, 2025a.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025b.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 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*, pp. 21807–21818, 2024.

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*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

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.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-harness: End-to-end optimization of model harnesses. *arXiv preprint arXiv:2603.28052*, 2026.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet ++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pp. 129–147. Springer, 2025b.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024.

Jiahang Lin, Shichun Liu, Chengjun Pan, Lizhi Lin, Shihan Dou, Xuanjing Huang, Hang Yan, Zhenhua Han, and Tao Gui. Agentic harness engineering: Observability-driven automatic evolution of coding-agent harnesses. *arXiv preprint arXiv:2604.25850*, 2026.

Feng-Lin Liu, Hongbo Fu, Xintao Wang, Weicai Ye, Pengfei Wan, Di Zhang, and Lin Gao. Sketchvideo: Sketch-based video generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23379–23390, 2025a.

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

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Midjourney. Midjourney. `https://www.midjourney.com`, 2026.

OpenAI. Sora2.0. `https://openai.com/index/sora-2/`, 2025.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, et al. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, pp. 4195–4205, 2023.

Bowen Ping, Chengyou Jia, Minnan Luo, Changliang Xia, Xin Shen, Zhuohang Dang, and Hangwei Qian. Paco-rl: Advancing reinforcement learning for consistent image generation with pairwise reward modeling. *arXiv preprint arXiv:2512.04784*, 2025.

Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-factory: A unified framework for reinforcement learning in flow-matching models. *arXiv preprint arXiv:2602.12529*, 2026. URL `https://arxiv.org/abs/2602.12529`.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Neural Information Processing Systems*, 36:53728–53741, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

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

Uriel Singer, Amit Zohar, Yuval Kirstain, Shelly Sheynin, Adam Polyak, Devi Parikh, and Yaniv Taigman. Video editing via factorized diffusion distillation. In *European Conference on Computer Vision*, pp. 450–466. Springer, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

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

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025b.

Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379*, 2025c.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

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

Hao Yang, Zhiyu Tan, Jia Gong, Luozheng Qin, Hesen Chen, Xiaomeng Yang, Yuqing Sun, Yuetan Lin, Mengping Yang, and Hao Li. Omni-video 2: Scaling mllm-conditioned diffusion for unified video generation and editing. *arXiv preprint arXiv:2602.08820*, 2026.

Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. Dc-controlnet: Decoupling inter- and intra-element conditions in image generation with diffusion models. *International Conference on Computer Vision*, 2025a.

Hongji Yang, Yucheng Zhou, Wencheng Han, Runzhou Tao, Zhongying Qiu, Jianfei Yang, and Jianbing Shen. Hicogen: Hierarchical compositional text-to-image generation in diffusion models via reinforcement learning. *arXiv preprint arXiv:2511.19965*, 2025b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, pp. 3836–3847, 2023.

Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A THEORETICAL ANALYSIS

This appendix provides the proofs of Lemma 1, Lemma 2, Proposition 1, and Corollary 1 stated in the main text. Appendix A.1 presents the stage-wise surrogate optimization view of the two-stage training and proves the stationarity of the frozen CogDirector criterion. Appendix A.2 proves the noise-concentration property of the weighted evaluator harness, the misranking and selection-regret bounds of best-of-N selection, and the corollary on lower-noise adaptive evaluation.

A.1 STAGE-WISE SURROGATE OPTIMIZATION

Population objectives. The factorization in Eq. 1 motivates an ideal population-level objective

$$J(\phi, \theta) = \mathbb{E}_{\substack{\mathcal{C}\sim\mathcal{D},\ \mathcal{R}\sim\pi_\phi(\cdot|\mathcal{C}),\\ \mathcal{V}\sim p_\theta(\cdot|\mathcal{R},\mathcal{C})}} \left[Q(\mathcal{V}; \mathcal{R}, \mathcal{C})\right], \tag{18}$$

In practice, evaluating the downstream generation quality for every sampled reasoning output is prohibitively expensive for video generation. CogOmniControl therefore uses two task-aligned surrogate objectives. For the reasoning stage, we define

$$J_{\mathrm{R}}(\phi) = \mathbb{E}_{\mathcal{C}\sim\mathcal{D},\ \mathcal{R}\sim\pi_\phi(\cdot|\mathcal{C})} \left[R_{\mathrm{CogDirector}}(\mathcal{R}, \mathcal{C})\right], \tag{19}$$

where $R_{\mathrm{CogDirector}}$ combines the holistic and factual rewards defined in Sec. 3.2. Using the score-function identity, its population gradient can be written as

$$\nabla_\phi J_{\mathrm{R}}(\phi) = \mathbb{E}_{\mathcal{R}\sim\pi_\phi(\cdot|\mathcal{C})} \left[R_{\mathrm{CogDirector}}(\mathcal{R}, \mathcal{C}) \nabla_\phi \log \pi_\phi(\mathcal{R}|\mathcal{C})\right]. \tag{20}$$

Our GRPO implementation constructs a sample-based clipped surrogate using group-relative advantages. Accordingly, J_{R} should be interpreted as the population objective underlying the reasoning-stage optimization rather than as an objective for which every finite-sample GRPO update is guaranteed to produce monotonic improvement. After this stage, CogDirector is frozen at ϕ^*. The generator-stage population objective is

$$J_{\mathrm{G}}(\theta; \phi^*) = \mathbb{E}_{\substack{\mathcal{C}\sim\mathcal{D},\ \mathcal{R}\sim\pi_{\phi^*}(\cdot|\mathcal{C}),\\ \mathcal{V}\sim p_\theta(\cdot|\mathcal{R},\mathcal{C})}} \left[R_{\mathrm{visual}}(\mathcal{V}, \mathcal{R}, \mathcal{C})\right]. \tag{21}$$

More generally, let τ denote the stochastic denoising trajectory that produces the final video $\mathcal{V}(\tau)$. The corresponding score-function form is

$$\nabla_\theta J_{\mathrm{G}} = \mathbb{E}\left[R_{\mathrm{visual}}(\mathcal{V}(\tau), \mathcal{R}, \mathcal{C}) \nabla_\theta \log p_\theta(\tau|\mathcal{R}, \mathcal{C})\right], \tag{22}$$

which follows from the stochastic-differential-equation view of RL for flow-matching models (Liu et al., 2025b; Xue et al., 2025). The implemented RFT objective provides a sample-based surrogate for this population objective.

Proof of Lemma 1. During generator RFT, CogDirector is fixed at ϕ^*. Therefore, the conditional reasoning distribution $\pi_{\phi^*}(\mathcal{R} \mid \mathcal{C})$ does not depend on the generator parameters θ. Likewise, the semantic representation produced by CogDirector and the CogDirector-based reward function $R_{\mathrm{visual}}^{\phi^*}$ have fixed parameters throughout generator optimization.

Let τ denote the stochastic denoising trajectory and $\mathcal{V}(\tau)$ its resulting video. The generator-stage objective in Eq. 21 can be written as $J_{\mathrm{G}}(\theta; \phi^*) = \mathbb{E}\big[R_{\mathrm{visual}}^{\phi^*}(\mathcal{V}(\tau), \mathcal{R}, \mathcal{C})\big]$, where only the trajectory distribution $p_\theta(\tau \mid \mathcal{R}, \mathcal{C})$ changes with θ, with score-function form given in Eq. 22. Hence updating θ changes the generator distribution but not the CogDirector conditioning distribution or the parameters of the reward function. If CogDirector were jointly updated, both quantities could change during generator optimization, creating a moving conditioning and evaluation target. This establishes the claimed stationarity with respect to generator updates.

The result concerns stationarity of the optimization criterion and does not imply monotonic improvement or convergence of finite-sample GRPO updates. Importantly, stationarity here refers to the parameterization of the conditioning and reward functional; the numerical reward assigned to a generated sample may change as p_θ changes. □

This analysis motivates the stage-wise training design without requiring a monotonic-improvement or convergence guarantee for the finite-sample clipped GRPO updates.

A.2 Adaptive Evaluator Selection

Setup. Let $\{\mathcal{V}_1, \ldots, \mathcal{V}_N\}$ denote the sampled candidates. For a fixed harness $\mathcal{H} = \{(e_k, w_k)\}_{k=1}^K$, let q_{ik} denote the latent quality of candidate i under criterion k, and let the observed evaluator score satisfy

$$e_{ik} = q_{ik} + \xi_{ik}.$$

Define

$$Q_i^{\mathcal{H}} = \mathbf{w}^\top \mathbf{q}_i, \qquad S_i = \mathbf{w}^\top \mathbf{e}_i = Q_i^{\mathcal{H}} + \mathbf{w}^\top \boldsymbol{\xi}_i,$$

where $\mathbf{q}_i = (q_{i1}, \ldots, q_{iK})^\top$, $\mathbf{e}_i = (e_{i1}, \ldots, e_{iK})^\top$, and $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^\top$. Hence $\varepsilon_i = \mathbf{w}^\top \boldsymbol{\xi}_i$ denotes the aggregate evaluation error, consistent with Assumption 1. Let $i^* = \arg\max_i Q_i^{\mathcal{H}}$ and define

$$\Delta = Q_{i^*}^{\mathcal{H}} - \max_{j \neq i^*} Q_j^{\mathcal{H}} > 0, \qquad B = Q_{i^*}^{\mathcal{H}} - \min_j Q_j^{\mathcal{H}}, \tag{23}$$

consistent with Eq. 13 in Proposition 1. The score noises are independent across sampled candidates; by Lemma 2, each ε_i is sub-Gaussian with variance proxy $\sigma_{\mathcal{H}}^2 = \mathbf{w}^\top \Sigma_{\mathcal{H}} \mathbf{w}$.

Proof of Lemma 2. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda \varepsilon_i)\right] &= \mathbb{E}\left[\exp\left(\lambda \mathbf{w}^\top \boldsymbol{\xi}_i\right)\right] \\ &\leq \exp\left(\frac{\lambda^2}{2} \mathbf{w}^\top \Sigma_{\mathcal{H}} \mathbf{w}\right).\end{aligned}$$

Hence ε_i is sub-Gaussian with variance proxy

$$\sigma_{\mathcal{H}}^2 = \mathbf{w}^\top \Sigma_{\mathcal{H}} \mathbf{w}.$$

When the evaluator errors are uncorrelated, $\Sigma_{\mathcal{H}}$ is diagonal with entries $\{\sigma_k^2\}_{k=1}^K$, yielding

$$\sigma_{\mathcal{H}}^2 = \sum_{k=1}^K w_k^2 \sigma_k^2.$$

□

Proof of Proposition 1. For any suboptimal candidate $j \neq i^*$, a misranking can occur only if $S_j \geq S_{i^*}$. Since

$$S_j - S_{i^*} = (Q_j^{\mathcal{H}} - Q_{i^*}^{\mathcal{H}}) + (\varepsilon_j - \varepsilon_{i^*})$$

and $Q_{i^*}^{\mathcal{H}} - Q_j^{\mathcal{H}} \geq \Delta$, we obtain

$$S_j \geq S_{i^*} \quad \Longrightarrow \quad \varepsilon_j - \varepsilon_{i^*} \geq \Delta.$$

By the assumed sub-Gaussian tail bound,

$$\mathbb{P}\left[S_j \geq S_{i^*}\right] \leq \exp\left(-\frac{\Delta^2}{2\nu_{\mathcal{H}}^2}\right).$$

Applying the union bound over the $N-1$ suboptimal candidates gives

$$\mathbb{P}[\text{misrank}] \leq (N-1) \exp\left(-\frac{\Delta^2}{2\nu_{\mathcal{H}}^2}\right),$$

which establishes Eq. 14.

For the regret, define $L = Q_{i^*}^{\mathcal{H}} - Q^{\mathcal{H}}(\mathcal{V}^*)$. If the best candidate is selected, $L = 0$; otherwise $0 < L \leq B$. Therefore,

$$\begin{aligned}\mathbb{E}[L] &= \mathbb{E}\left[L \mathbf{1}\{\text{misrank}\}\right] \\ &\leq B\, \mathbb{P}[\text{misrank}] \\ &\leq B(N-1) \exp\left(-\frac{\Delta^2}{2\nu_{\mathcal{H}}^2}\right),\end{aligned}$$

which establishes Eq. 15.

If the candidate-level errors are independent and each is sub-Gaussian with proxy $\sigma_{\mathcal{H}}^2$, then $\varepsilon_j - \varepsilon_{i^*}$ is sub-Gaussian with proxy at most $2\sigma_{\mathcal{H}}^2$. Setting $\nu_{\mathcal{H}}^2 = 2\sigma_{\mathcal{H}}^2$ recovers Eqs. 16 and 17. □

Proof of Corollary 1. For a fixed candidate set, both bounds in Proposition 1 are monotonically increasing in $\nu_{\mathcal{H}}^2$. Therefore, if $\nu_{\text{adaptive}}^2 < \nu_{\text{fixed}}^2$, substituting the two variance proxies into the corresponding bounds immediately gives strictly tighter misranking and expected-regret bounds for the adaptive harness. □

Remark 1 (Scaling with the number of candidates). With an oracle selector, the expected best-of-N quality is non-decreasing in N, since increasing the number of samples enlarges the candidate set. With a noisy evaluator, however, the quality of the actually selected candidate additionally depends on the effective score-noise level $\nu_{\mathcal{H}}$ and the quality gap Δ_N between the top candidates. Therefore, Proposition 1 does not imply that increasing N alone always improves the selected quality. Instead, test-time scaling benefits from both a stronger candidate pool and a sufficiently reliable evaluator, which motivates combining best-of-N sampling with the input-adaptive harness.

B Full Related Work

Video Generation. With the rapid development of image (Rombach et al., 2022; Podell et al., 2023; Peebles & Xie, 2023; Labs, 2024) and video (Hong et al., 2022; Yang et al., 2024; HaCohen et al., 2024; Kong et al., 2024; Wan et al., 2025) generative models, diffusion models have been proven to produce high-fidelity visual content and are widely applied in diverse domains, including artistic creation, animation production, visual special effects and game development (Brooks et al., 2024; Midjourney, 2026). To faithfully realize specific creative intentions, conditional guidance has evolved from abstract natural language to diverse explicit constraints for precise control. Early breakthroughs introduce additional adapters (Zhang et al., 2023; Ye et al., 2023; Li et al., 2025b; Yang et al., 2025a; Guo et al., 2023; Zhao et al., 2023; Jiang et al., 2025; Guo et al., 2024; Lin et al., 2024; Liu et al., 2025a) to support condition injection without compromising the original generative quality. However, these adapter-based paradigms often exhibit limited flexibility in handling diverse conditions, particularly in those that are non-pixel-aligned or serve merely as visual references. To achieve omni-level generation, OmniGen (Xiao et al., 2025) and OmniGen2 (Wu et al., 2025a) integrated autoregressive transformers with diffusion to realize unified generation. OmniControl (Tan et al., 2025) and UNO (Wu et al., 2025b) introduced in-content visual generation.

Omni-level generation has also been extended into the video domain, the emergence of proprietary models, such as Seedance2.0 (Seedance et al., 2026), Kling-O1 (Team et al., 2025), Sora2 (OpenAI, 2025), Vidu (AI, 2026), and Veo3 (Google, 2025), has established a transformative vision for omni-level video generation. However, current open-source models still fail to realize robust unified video generation. VACE (Jiang et al., 2025), UniVideo (Wei et al., 2025), and VINO (Chen et al., 2026) attempted to achieve omni-level generation by integrating various basic tasks, they often lack a deep understanding across diverse conditions. In contrast, OmniWeaving (Pan et al., 2026) successfully incorporated the abstract reasoning of VLM into the video diffusion model to execute complex multimodal compositional tasks. However, the reasoning processes of its LLM components have not yet undergone professional evaluation or systematic benchmarking on creative intentions, leaving the model without sufficient guidance when tackling more challenging tasks.

Reinforcement Learning for Visual Generation. Inspired by the success of LLM fine-tuning using RL from human feedback, RL for visual generation is gaining momentum. For example, DDPO (Black et al., 2024), Diffusion-DPO (Wallace et al., 2024) and DPOK (Fan et al., 2023) introduced Direct Preference Optimization (Rafailov et al., 2023) into T2I Diffusion to align with human preference. Motivated by DeepSeek-R1 (Guo et al., 2025) using GRPO (Shao et al., 2024) to provide denser rewards through computing relative rewards in a sample group, Flow-GRPO (Liu et al., 2025b) and DanceGRPO (Xue et al., 2025) extended this paradigm into flow-matching models (Liu et al., 2022) by transforming the deterministic ODE formulation into a stochastic SDE, thereby enabling effective online exploration and policy alignment. Beyond this, several GRPO-based studies (Wang et al., 2025; Li et al., 2025a; He et al., 2025b; Yang et al., 2025b) have focused on refining reward design Liu et al. (2026); Wu et al. (2025c) to enhance performance in visual generation.

Harness Engineering and Test-Time Scaling. For language models, lm-evaluation-harness (Gao et al., 2024) develops a unified harness for generic LLM evaluation, and recent studies (Lee et al., 2026; Lin et al., 2026) further show that an external harness can plan and orchestrate tools to scale the capability of a fixed model at inference time. In visual generation, VLM-as-a-Judge (Zheng et al.,

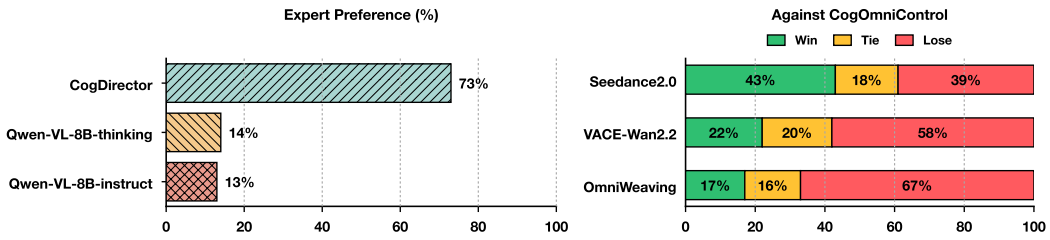

Figure 6: **Left:** The expert preference of output from different VLMs; **Right:** The human evaluation results of the generated videos.

2023; Chen et al., 2024) and dimension-wise benchmarks such as VBench (Huang et al., 2024) provide reliable feedback for test-time scaling over sampled candidates. However, existing video selection protocols apply a *fixed* set of evaluators to all samples, ignoring that different conditions carry distinct creative intents and that the evaluators should therefore be selected per input. In contrast, CogDirector emits both the reasoning and the evaluator set in a single forward pass, adaptively assigning the evaluators conditioned on the input.

B.1 Human Studies

We collect expert preferences from 10 expert annotators across 100 samples under CogDirector, Qwen-VL-8B-Instruct, and Qwen-VL-8B-Thinking. These experts are students specializing in filmmaking and directing at various high education institutions. As shown in Fig. 6, CogDirector received 73% preference from human evaluators. The Kendall's $\mathcal{W}$ among the evaluators is 0.87, indicating a high degree of agreement among them. Furthermore, we conduct a human preference study comparing the generated results of CogOmniControl with Seedance2.0 and OmniWeaving. This human subject study also invited 10 experts with filmmaking and directing backgrounds to participate. As shown in Fig. 6, CogOmniControl and Seedance2.0 are competitive, while CogOmniControl substantially outperforms OmniWeaving and VACE-Wan2.2.

C Training Details

C.1 Experiment Setup

Training Data. Our training corpus is constructed from the same in-house professional animation production pipeline that provides the benchmark, together with general controllable generation data collected from the community and public sources. All clips appearing in CogControlBench and CogReasonBench are excluded from the training set to avoid any leakage. In total, we collect 20k condition–solution pairs for the SFT of CogDirector, 1.2k condition sets for its RFT, and 300k condition–video pairs for the training of the generator, in which the professional workflow data accounts for 32%.

Implementation Details. Experiments are conducted on the Qwen3-VL-8B-Thinking (Bai et al., 2025) as the base VLM and Wan2.2-T2V-14B (Wan et al., 2025) as the base DiT with 32 NVIDIA H20 96GB GPUs. For SFT in CogDirector, we employ LoRA (Hu et al., 2022) training with a rank of 16 and an alpha of 64. The SFT is performed for 3 epochs with a learning rate of 1e-5. For RFT in CogDirector, we train our model with an initial learning rate of 1e-6 for 500 steps. For SFT in DiT, we implement a three-stage training strategy using LoRA with a rank of 256. In stage-1, we train only LoRA for in-context generation. In stage 2, we introduce freezing CogDirector and a trainable connector. Finally, we perform joint training of the LoRA and connector. The connector is set to two-layer MLPs with GELU. The w_k and λ in Eq. 4 are set to 1, respectively. The rollout number in GRPO is set to 4.

C.2 Evaluation Metrics

To comprehensively evaluate the performance of CogOmniControl in controllable video generation, we utilize numeric metrics based on VBench (Huang et al., 2024) and a VLM-as-a-Judge (Zheng et al., 2023) paradigm, employing Gemini 3.1-Pro (Google) as the authoritative evaluator. Our evaluation focuses on two dimensions:

Table 5: Training Setting of DiT.

	SFT Stage-1	SFT Stage-2	SFT Stage-3	RFT
Hyperparameter				
Trainable module	LoRA	Connector	LoRA+Connector	LoRA+Connector
CogDirector included	no	yes	yes	yes
Learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
Learning rate scheduler	Constant	Constant	Constant	Cosine
Weight Decay	3×10^{-2}	3×10^{-2}	3×10^{-2}	0
Gradient Norm	5×10^{-2}	5×10^{-2}	5×10^{-2}	1.0
GPUs	$32 \times$ H20s	$32 \times$ H20s	$32 \times$ H20s	$16 \times$ H20s
Training steps	10k	500	10k	200
Resolution	480P/720P	720P	720P	256P
Batch Size	8	4	4	8

Condition Following. The core of our evaluation lies in whether CogOmniControl faithfully adheres to the creative intent implied by the condition set $\{V_{ctrl}, I_{ref}, T_{desc}\}$. Unlike traditional methods that treat conditions as isolated constraints, we assess the model's ability to interpret these multimodal signals as a holistic objective. For this task, the evaluation of multimodal intent alignment is based on the following considerations: whether the model effectively resolves conflicts between disparate conditions, whether it integrates conditions accurately when significant discrepancies exist, and whether it can infer plausible physical properties or dynamic effects based on the association among the conditions. Besides, the evaluation also includes the preservation of the visual information from I_{ref} and the instruction following from T_{desc} in this task.

Visual Quality. Visual quality evaluates the aesthetic quality, imaging quality, temporal flickering, motion smoothness and dynamic degree of the generated video inspired by VBench (Huang et al., 2024). Besides, this type of evaluation also provides dimensions on identity consistency and dynamic plausibility.

In Table 3, scores originally on a [0-5] scale are normalized to [0-1] when computing the average; the final average score is presented on the 0-1 scale. We also report the results with test-time scaling, which outperforms the open-source models.

C.3 Training Settings of Video Generator

The training details of CogOmniControl are shown in Tab 5. The three stages of SFT are training for 1) bringing in-context generation ability; 2) aligning CogDirector into DiT; 3) joint training. For RFT, we follow flow-factory (Ping et al., 2026) and PaCo-RL (Ping et al., 2025) to perform GRPO training in low resolution (256P) and inference in higher resolution (720P).

D Evaluator Harness

In Tab 6, we report the types and times of calling the evaluator for test-time scaling. The common evaluator, like `[Artifact Detector]`, `[Prompt Following]`, `[Temporal Smoothness]` will be called all the time, and other specific evaluators are called based on the reasoning. For example, if the input condition is a storyboard, CogDirector identifies whether there are handwritten annotations that must be followed. After generation, it then invokes the `[Storyboard Annotation Following]` evaluator to verify whether the video adheres to those specific instructions. We provide the examples of evaluators as follows:

Table 6: The frequency of tools in CogOmniControl used in CogControlBench.

#	Evaluator	Times
1	Artifact Detector	200
2	Prompt Following	200
3	Temporal Smoothness	200
4	Control Video Semantic Consistency	136
5	Interaction Logic	132
6	Ref Image Pixel Consistency	17
7	Ref Image Visual Consistency	106
8	Reference Style Consistency	156
9	Motion Smoothness	92
10	ID Consistency	106
11	Cross-modal Causality	117
12	Physical Dynamic	49
13	Storyboard Annotation Following	8

Artifact Detector

Role
You are a professional AI Video Quality Inspection Expert. Your responsibility is to identify common AI-generated artifacts and defects in videos. Focus on detecting AI-specific issues such as multiple heads/limbs, deformation, floating, rendering failures, abnormal noise, smearing marks, etc.

Inputs
You will receive the following four inputs:
Reference Image (Ref Image) | Image | The reference image provided by the user.
Control Video (Control Video) | Video | Control signals such as Pose/Depth/Line Art/Storyboard.
Text Prompt (Prompt) | Text | The generation requirements described by the user.
Generated Video (Generated Video) | Video | The AI-generated video to be evaluated.

Rules
1. Multiple Head/Limb Detection: Detect if the same character or object has extra heads, limbs, fingers, etc.
2. Deformation Detection: Detect non-physical twisting, stretching, or crushing of objects/characters.
3. Floating Detection: Detect if objects violate gravity by floating or appearing unstable/ungrounded.
4. Rendering Failure Detection: Detect rendering issues such as local blurring, smearing, or noise accumulation.
5. Identity Distortion: Detect facial distortions, blurring, or abnormalities.
6. Background Collapse: Detect unnatural distortion, noise, or blurring in the background.

Common AI Artifact Types (For Reference)
Multi-head/Polycephaly: The same person has two or more heads.
Extra Fingers: Extra fingers on hands or webbed/fused fingers.
Limb Entanglement: Limbs twisted, knotted, or detached from the body.
Facial Deformation: Facial distortion, blurring, abnormal eyes/mouth.
Object Floating: Objects suspended in the air against gravity.
Hovering Contact: Objects contact the ground but look like they are stepping on cotton.
Smearing Marks: Visible blurred "smearing" sensation in local areas.
Abnormal Noise: Significantly more noise in specific areas compared to others.
Background Collapse: Background distortion, noise, or objects disappearing for no reason.

5-Point Scoring Criteria (0-5)

5 - No Artifacts: The video is completely free of any AI-generated artifacts. The frame is clean and natural; limbs are correct; objects are stable; rendering is perfect. No issues with multiple heads/limbs, deformation, floating, or abnormal noise.

4 - Basically Clean: The video is mostly free of obvious artifacts. There are 1-2 extremely subtle imperfections, such as minor noise or very slight local blurring, which do not affect the overall viewing experience and are not typical AI artifacts.

3 - Slight Artifacts: The video contains some of the following: 1-2 instances of slight extra fingers/limbs; 1-2 instances of slight facial deformation; minor abnormal noise or local blurring; 1 instance of slight object floating; 1 instance of local smearing.

2 - Moderate Artifacts: The video contains multiple artifacts: Frequent extra fingers/limbs; Obvious facial deformation or blurring; Multiple floating objects; Multiple areas of abnormal noise; Multiple smearing marks; Background collapse.

1 - Severe Artifacts: The video contains severe artifacts. Multiple typical AI generation issues are clearly visible, such as persistent multiple heads/limbs, severe facial distortion, or large-scale rendering failures, seriously impacting the viewing experience.

0 - Total Collapse: The video is riddled with AI artifacts. Multiple heads/limbs appear persistently; faces are completely unrecognizable; objects totally violate physical laws; rendering has completely failed; the content is nearly unidentifiable.

Output Format
Please output strictly in the following JSON format without any other content:

```
"""
json
{
  "evaluator": "Artifact Detector",
  "score": <0-5>,
  "findings": "detail findings",
  "summary": "summary"
}
"""
```

Storyboard Annotation Evaluators

Role
You are a professional Storyboard Annotation Implementation Auditor. Your responsibility is to determine whether the generated video faithfully follows the text annotation instructions attached to the Control Video or Storyboard (e.g., "swaying shadows of trees," "smiling," "wind rising," etc.). These annotations are independent of the user's original prompt and represent additional creative intent from the director or planner that must be correctly executed.

Inputs
You will receive the following four inputs:
Reference Image (Ref Image) | Image | The reference image provided by the user.
Control Video (Control Video) | Video | Control signals such as Pose/Depth/Line Art/Storyboard.
Text Prompt (Prompt) | Text | The generation requirements described by the user.
Generated Video (Generated Video) | Video | The AI-generated video to be evaluated.

Rules
1. Prioritize Annotation Identification: First, identify all text annotation content within the Control Video/Storyboard.
2. Point-by-Point Verification: Check each text annotation individually to see if it is correctly rendered in the video.
3. Dynamic Assessment: Descriptions of actions/dynamics in the annotations must exhibit clear motion; static frames are unacceptable.
4. Annotation Independence: Annotations may exist independently of the Prompt and Ref Image as distinct creative commands.
5. Temporal Matching: The timing or duration described in the annotations must match the corresponding moments in the video.

Annotation Types & Evaluation Points
1. Scene Dynamics Examples: "Swaying tree shadows," "clouds drifting by," "shimmering water surface."
Evaluation Points: Whether background elements show clear motion and if the movement conforms to natural laws.
2. Character Actions
Examples: "Smiling," "turning around," "waving," "blinking."
Evaluation Points: Whether the character performed the specified action and if the movement is natural and fluid.
3. Environmental Atmosphere
Examples: "Wind rising," "rain falling," "falling leaves," "flickering candlelight."
Evaluation Points: Whether environmental dynamics match the annotations and create the intended atmosphere.
4. Emotional Expression
Examples: "Sad eyes," "corners of the mouth turned up," "furrowed brows."
Evaluation Points: Whether the character's facial expressions match the annotations and accurately convey the emotion.
5. Object Dynamics
Examples: "Flag fluttering," "curtains swaying," "sparks flying."
Evaluation Points: Whether object motion is distinct and physically plausible.

5-Point Scoring Criteria (0-5)
5 - Perfect Adherence: All storyboard annotations are perfectly executed. Dynamic effects are significant and natural, timing is precise, and every detail described in the annotations is faithfully presented.
4 - Basic Adherence: Storyboard annotations are largely executed. There are 1-2 minor deficiencies (e.g., slightly small range of motion or slight timing deviation), but the core annotation content is correctly presented without affecting the overall creative intent.
3 - Partial Adherence: Storyboard annotations are partially executed, with the following issues: 1-2 annotations not fully executed; Dynamic effects are indistinct (too subtle); 1 significant timing deviation; Misinterpretation of 1 annotation.
2 - Substantial Omission: Significant failure to follow storyboard annotations, involving multiple issues: Multiple annotations not executed; Extensive lack of dynamic effects or extremely weak motion; Multiple timing deviations; Obvious errors in understanding annotations.
1 - Severe Omission: Storyboard annotations are almost entirely ignored. Only a tiny fraction of annotations show weak implementation; most are neglected or completely misunderstood, leading to a severe loss of creative intent.
0 - Complete Non-compliance: None of the text annotations on the storyboard are executed. The video fails to present any annotation content, or the annotations are executed with total error (e.g., annotation says "smiling" but the character is weeping).

Output Format
Please output strictly in the following JSON format without any other content:

```
"""
json
{
  "evaluator": "Storyboard Annotation Evaluators",
  "score": <0-5>,
  "findings": "detail findings",
  "summary": "summary"
}
"""
```

Cross- modal Causality

Role
You are a professional Cross-modal Causal Reasoning Analysis Expert. Your responsibility is to judge whether a video correctly "interpolates" the causal relationships implied between multi-modal inputs (Text + Image + Control Video). Key point: The combination of events/states in the text, existing elements in the image, and action cues in the control video should produce logically sound causal interactions.

Inputs
You will receive the following four inputs:
Reference Image (Ref Image) | Image | The reference image provided by the user.
Control Video (Control Video) | Video | Control signals such as Pose/Depth/Line Art/Storyboard.
Text Prompt (Prompt) | Text | The generation requirements described by the user.
Generated Video (Generated Video) | Video | The AI-generated video to be evaluated.

Rules
1. Causal Implication Identification: Identify causal relationships implied across the text, image, and control video (e.g., "Rain" + "Puddles in image: There should be ripples).
2. Interaction Effect Verification: Check if the video generates the appropriate causal interaction effects.
3. Causal Chain Integrity: Check if the causal relationship is complete (Cause $\rightarrow$ Process $\rightarrow$ Effect).
4. Reasonable Inference: The model is allowed to make reasonable inferences, but they must align with common sense.
5. No Hallucinated Causes: Inferences must be based on input information; do not create causalities that are completely unhinted at in the inputs.

Common Causal Patterns (For Reference)
Text "Rain" + Image has puddles $\rightarrow$ Ripples on water surface, wet ground.
Text "Hitting" + Control video has punching motion $\rightarrow$ Target should show dynamic impact effects

5-Point Scoring Criteria (0-5)
5 - Perfect Causal Interaction: All causal relationships implied by multi-modal inputs are correctly identified and perfectly presented. The causal chain is complete. Interaction effects are natural, logical, and meet or exceed expectations.
4 - Basically Correct: Causal relationships are largely correct with 1-2 minor flaws, such as slightly dull effects or minor timing deviations in interaction. However, the core causal chain is complete and does not affect overall logical consistency.
3 - Partially Missing: Causal relationships are partially correct, with issues such as: 1-2 key causal effects missing; Incomplete causal chain (cause exists without effect); One instance of unreasonable causal inference.
2 - Significant Deficiencies: Multiple causal issues present: Multiple causal effects failed to render; Causal timing is chaotic; Obvious unreasonable "hallucinated" causalities; Most expected interaction effects are ignored.
1 - Severe Lack of Causality: Almost all causal cues in the multi-modal inputs were ignored. The video lacks interaction effects entirely; elements exist in isolation without logical connection.
0 - Zero Causality: Elements in the video are completely isolated. No causal logic is utilized from the text/image/control video. The video appears as a collection of unrelated static or dynamic frames.

Output Format
Please output strictly in the following JSON format without any other content:
"""
json
{
  "evaluator": "Cross-modal Causality Evaluators",
  "score": <0-5>,
  "findings": "detail findings",
  "summary": "summary"
}
"""