FLOWCRED: HIERARCHICAL CREDIT ASSIGNMENT AND FUNCTIONAL COORDINATION FOR JOINT TEXT-TO-VIDEO POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt optimization and generator post-training are two complementary ways to improve text-to-video (T2V) generation, and recent systems increasingly train both modules jointly. However, optimizing a prompt policy and a video generator from the same terminal reward creates two ambiguities: prompt quality is confounded with stochastic generation, and the two modules may realize overlapping local improvements. We introduce FLOWCRED, a two-level framework for joint T2V post-training. First, hierarchical credit assignment uses matched prompt–seed rollout grids to separate prompt-level credit from prompt-conditioned generator residuals under shared randomness. Second, functional coordination probes alternative current-policy prompts at states from actual generation trajectories, providing generator-aware evidence for prompt optimization while selectively attenuating generator updates already supported by conditioning changes. Across multiple T2V benchmarks, FLOWCRED consistently improves matched joint-training baselines. On LTX-2.5, FLOWCRED improves a strictly matched joint-training control from 61.47 to 64.46 on T2V-CompBench and from 55.58 to 58.08 on StoryEval, without increasing decoded-video or reward-model-call budgets. Factorial and realized-update analyses show complementary contributions from prompt shaping and generator routing, while human evaluation and transfer across backbones and updater families provide further supporting evidence.

1 INTRODUCTION

Text-to-video (T2V) systems increasingly expose two trainable adaptation channels: a language-side prompt policy and the video generator itself. Prompt-side methods improve conditioning while leaving the visual model fixed (Hao et al., 2023; Cao et al., 2023; Ji et al., 2025; Cheng et al., 2025); generator post-training instead changes the diffusion/flow model using reward or preference feedback (Black et al., 2024; Fan et al., 2023; Wallace et al., 2024; Liu et al., 2025a; Xue et al., 2025; 2026). Recent systems optimize both modules jointly (Wang et al., 2026a; Cheng et al., 2026; Liu et al., 2026; Jia et al., 2026). Prompt changes can reorganize entities, relations, and event descriptions before generation, whereas generator updates address residual motion, appearance, and execution errors that conditioning alone cannot remove. When only one module is updated, a terminal reward can be treated as supervision for that module; once both are trainable, the same scalar signal must implicitly serve both attribution and coordination roles. Joint post-training is therefore attractive; once both channels are trainable, however, one terminal score leaves two distinct questions unresolved.

The first is *statistical credit*. A terminal reward conflates prompt quality with downstream generator randomness: identical prompts can score differently across seeds, while favorable trajectories can mask weaker prompt choices. The reward alone therefore does not identify which variation belongs to prompt choice and which remains after the prompt is fixed. For request u, prompt action p, stochastic generator trajectory $\mathcal{T}$, and terminal reward R, let $Q(u,p) = \mathbb{E}[R \mid u, p]$ and $V(u) = \mathbb{E}_p Q(u,p)$. Then

$$R - V(u) = \underbrace{Q(u,p) - V(u)}_{\text{prompt decision}} + \underbrace{R - Q(u,p)}_{\text{prompt-fixed trajectory}} . \tag{1}$$

This exposes two sources of terminal variation. The second question is *functional coordination*: even after statistical credit is separated, changing the prompt at a generator state may induce a response

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.

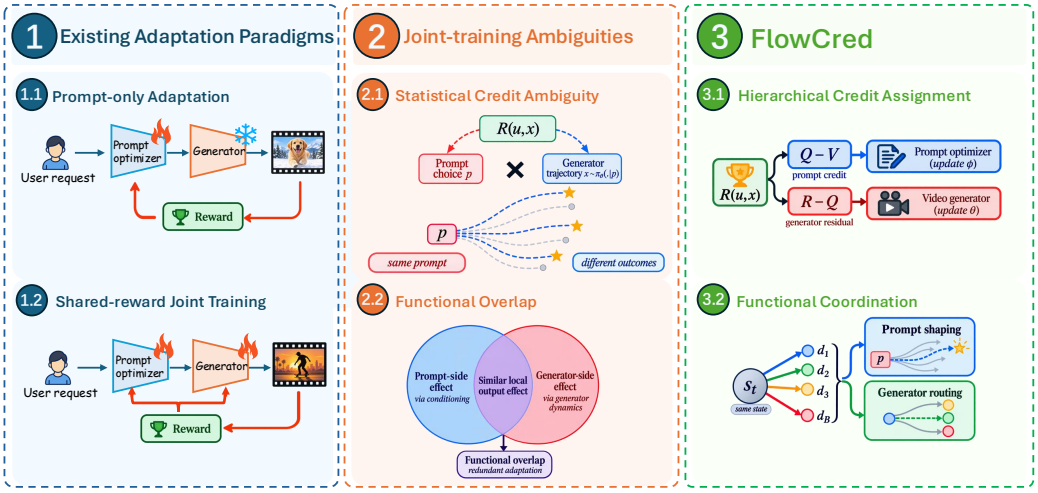


Figure 1: **Motivation and positioning of FLOWCRED.** The left panel contrasts prompt-only adaptation with shared-reward joint training. Once both modules are trained from a common terminal reward, two distinct issues arise: statistical credit ambiguity between prompt choice and prompt-fixed stochastic variation, and functional overlap between conditioning-side and generator-side local effects. FLOWCRED addresses them with hierarchical credit assignment ($Q - V$ versus $R - Q$) and functional coordination based on same-state response geometry.

aligned with the same reward-improving direction used by the downstream optimizer. The two modules can therefore duplicate part of the same local improvement. Statistical credit and functional overlap are thus distinct problems.

These two ambiguities require different remedies: variance reduction can make credit estimates cleaner without preventing functional overlap, while coordinating local updates does not by itself determine where terminal variation originated. Treating them separately lets joint post-training distinguish *who should receive credit* from *which module should realize an overlapping improvement*.

To address both, we introduce FLOWCRED, named for its joint treatment of credit assignment and coordination across the prompt-to-video generation flow. Figure 1 summarizes the two ambiguities and their corresponding mechanisms. **Hierarchical credit assignment** uses matched prompt–seed rollouts to separate prompt-level credit from prompt-conditioned generator residuals. **Functional coordination** probes alternative prompts at actual rollout states and uses same-state response geometry for generator-aware prompt shaping and reliability-gated generator routing.

The two mechanisms have different mathematical status: Eq. 1 is exact for any per-trajectory scalar score, whereas functional coordination is local and first-order. Its prompt term is auxiliary shaping and its generator rule is a minimum-overlap regularizer, not a conserved transfer of gradient mass. To keep the statistical claim separate from optimization heuristics, all affine calibration used to combine reward credit and functional prompt-credit terms is estimated once on disjoint training-only data and then frozen (Appendix D.3).

Our contributions are threefold.

(1) We introduce hierarchical credit assignment with matched prompt–seed rollouts, separating prompt-level credit from prompt-conditioned generator residuals under shared generator randomness. This gives the prompt policy and generator distinct estimators tied to their positions in the stochastic hierarchy.

(2) We develop functional coordination from same-state prompt responses, using local response geometry for generator-aware prompt shaping and reliability-gated routing of the generator reward update. It targets overlap that remains even after statistical credit has been separated.

(3) We evaluate FLOWCRED with controlled joint-training baselines, mechanism audits, repeated runs, human evaluation, and transfer across backbones and updater families. Across four T2V benchmarks, FLOWCRED consistently improves the matched controls without increasing decoded-video or reward-model-call budgets.

2 Related Work

Prompt optimization and generator post-training. Prompt-side methods optimize conditioning with a fixed generator (Hao et al., 2023; Cao et al., 2023; Ji et al., 2025; Cheng et al., 2025), whereas PromptRL, VANS, UniGRPO, and CAPE-T2V couple language- and generator-side adaptation in different ways (Wang et al., 2026a; Cheng et al., 2026; Liu et al., 2026; Jia et al., 2026). Generator alignment spans policy gradients, preference objectives, and differentiable reward optimization (Black et al., 2024; Fan et al., 2023; Wallace et al., 2024; Clark et al., 2024; Li et al., 2024; Xu et al., 2023; Kirstain et al., 2023; Liu et al., 2025a; Choi et al., 2026; Prabhudesai et al., 2024; Wang et al., 2026b); recent methods further refine reward shaping, trajectory allocation, and temporal credit (Xue et al., 2025; 2026; Luo et al., 2026; Zhu et al., 2026; Ye et al., 2026; Pu & Kong, 2026; Shen et al., 2026; Guan et al., 2026; Savani et al., 2026; Oba et al., 2026). These directions are complementary to our cross-module question. Modern open T2V systems span diverse backbone designs (Chen et al., 2024; Yang et al., 2025; Kong et al., 2024; Wan et al., 2025); our experiments hold the downstream harness fixed whenever numerical comparisons require matched conditions.

Credit assignment, variance reduction, and functional coordination. COMA counterfactual credit holds other agents fixed while marginalizing one action (Foerster et al., 2018), while leave-one-out baselines and common random numbers are standard tools for variance reduction (Ahmadian et al., 2024; Ng & Jordan, 2000). PCGrad/CAGrad manipulate conflicting parameter gradients across objectives (Yu et al., 2020; Liu et al., 2021). FlowCred instead separates statistical credit across the stochastic prompt-policy/generator hierarchy and measures functional overlap through same-state conditioning responses in the generator's native output space.

3 Method

3.1 Joint post-training: credit and coordination

Let $u \sim \mathcal{D}$ be a request, $p \sim \pi_\phi(\cdot \mid u)$ a prompt action, and $\mathcal{T} \sim \pi_\theta(\cdot \mid p)$ a stochastic generator trajectory ending in video x. The shared objective is

$$J(\phi, \theta) = \mathbb{E}_{u,p,\mathcal{T}}[R(u, x)]. \tag{2}$$

All rewards are evaluated against the original request u. We use *statistical credit* for how terminal variation is assigned among sampled decisions and *functional coordination* for how locally overlapping conditioning and generator changes are handled. Accordingly, hierarchical credit assigns terminal variation according to where it arises, while functional coordination acts only on local function changes that both trainable modules can express. We refer to π_ϕ as the *prompt optimizer* or *prompt policy*.

3.2 Hierarchical credit assignment

Define $Q_\theta(u, p) = \mathbb{E}_{\mathcal{T}\sim\pi_\theta(\cdot|p)}[R(u, x)]$ and $V_{\phi,\theta}(u) = \mathbb{E}_{p\sim\pi_\phi(\cdot|u)}[Q_\theta(u, p)]$; we suppress parameter subscripts below. Thus $R - V = (Q - V) + (R - Q)$ separates prompt choice from prompt-fixed stochasticity. For each request, sample K candidates $p_{1:K} \sim \pi_\phi(\cdot \mid u)$ and reuse the same M generator-randomness seeds $z_{1:M}$ for every candidate:

$$x_{k,m}, \mathcal{T}_{k,m} = G_\theta(p_k; z_m), \qquad R_{k,m} = R(u, x_{k,m}). \tag{3}$$

The matched rollout grid yields the paired leave-one-out estimators

$$\begin{aligned} \widehat{A}^{P,R}_k &= \frac{1}{M}\sum_m \left(R_{k,m} - \frac{1}{K-1}\sum_{j\neq k} R_{j,m} \right), \\ \widehat{A}^{G}_{k,m} &= R_{k,m} - \frac{1}{M-1}\sum_{m'\neq m} R_{k,m'}. \end{aligned} \tag{4}$$

The first contrast compares prompt decisions under common generator randomness and estimates $Q - V$; the second compares trajectories after the prompt is fixed and estimates $R - Q$. Positive

same-seed reward covariance lowers cross-prompt contrast variance (Appendix A). The prompt optimizer uses $\widehat{A}^{P,R}$, while the generator reward branch uses $\widehat{A}^{G}$. Reward-component calibration is estimated on disjoint training-only data and frozen, so no statistic from the current $K \times M$ grid enters $R_{k,m}$ (Appendix D.3).

Controlled global-LOO baseline. Matched Global-LOO uses the identical $K \times M$ grid but gives both modules one global leave-one-out trajectory advantage (Appendix A.2). Averaged over the M trajectories for a prompt, this control preserves the same coarse prompt ordering up to positive scale. The substantive statistical change in hierarchical credit is therefore downstream: generator credit becomes the prompt-conditional residual $R - Q$ rather than a global outcome residual.

3.3 Functional coordination via same-state responses

Hierarchical credit separates endpoint variation but does not describe whether prompt conditioning and the generator reward step realize similar local function changes. We measure that overlap at states actually encountered by the generator and reuse the same response geometry for prompt shaping and generator routing. An encountered state $s_t^{k,m}$ contains the noisy latent, timestep, and scheduler quantities required for one transition on trajectory (k, m). Let $\mu_t = \mu_\theta(s_t^{k,m}, p_k)$ be the backend-native transition output. For Full, the base updater's reward branch on this trajectory is weighted by $\widehat{A}^{G}_{k,m}$ from Eq. 4; stabilization/reference terms remain in a separate branch:

$$J_{\mathrm{base},t} = J_{\mathrm{reward},t} + J_{\mathrm{reg},t}, \qquad q_t^R = \nabla_{\mu_t} J_{\mathrm{reward},t}, \quad g_t^R = F_t^{-1} q_t^R. \tag{5}$$

Here F_t is the backend's local metric (the transition Fisher for Flow-GRPO), q_t^R is the reward covector, and g_t^R is its metric-mapped local ascent direction; we write $\|v\|_{F_t}^2 = v^\top F_t v$. Reference/KL regularization does not enter q_t^R and is restored unchanged after routing.

We cache the selected state rather than an alternative continuation. With the detached behavior snapshot $\theta^- = \mathrm{sg}(\theta)$, replacing only the executed prompt p_k by another current-policy candidate p_j gives

$$\begin{aligned} d_{j\leftarrow k,t} &= \mu_{\theta^-}(s_t^{k,m}, p_j) - \mu_{\theta^-}(s_t^{k,m}, p_k), \\ \bar{d}_{j\leftarrow k,t} &= \mathrm{TrustClip}_{F_t}(d_{j\leftarrow k,t}), \qquad c_{j\leftarrow k,t} = \bar{d}_{j\leftarrow k,t}^{\top} q_t^R. \end{aligned} \tag{6}$$

The probe is intentionally one-transition and same-state: $s_t^{k,m}$ is seed/state dependent, but within each comparison the state, stochastic draw, and behavior snapshot are fixed and only text conditioning changes. Continuing p_j would mix this local conditioning response with later nonlinear state evolution and additional stochasticity. We use $\mathrm{TrustClip}_{F_t}(d) = d \min\Big(1, \sqrt{\tau_{\mathrm{batch}}/(d^\top F_t d + \epsilon_{\mathrm{num}})}\Big)$, where $\epsilon_{\mathrm{num}} = 10^{-12}$ and τ_{batch} is the detached 95th percentile of the current-batch native-metric response energy $d^\top F_t d$. Thus in-region responses are unchanged and only unusually large conditioning shifts are rescaled. The score $c_{j\leftarrow k,t}$ is the resulting local directional derivative along an observed prompt response; its sign records agreement with the reward branch. Full-trajectory utility of prompt candidates is still measured by the $K \times M$ rollouts in Eq. 3.

Generator-aware prompt shaping. Let $\mathcal{S}_{k,m}$ denote the selected routed transition indices on trajectory (k, m) (one by default). Cross-fitting avoids scoring p_j on states produced by itself:

$$C_j = \frac{1}{K-1} \sum_{k \neq j} \mathbb{E}_{m,\, t \in \mathcal{S}_{k,m}}[c_{j\leftarrow k,t}], \qquad A_j^P = \mathrm{Cal}_R(\widehat{A}_j^{P,R}) + \lambda_F \, \mathrm{Cal}_C(C_j). \tag{7}$$

For $X \in \{R, C\}$, $\mathrm{Cal}_X(x) = (x - \mu_X^{\mathrm{cal}})/(\sigma_X^{\mathrm{cal}} + 10^{-6})$ uses statistics estimated once on a disjoint training-only calibration split and then frozen; Cal_R acts on terminal prompt credit and Cal_C on functional evidence. The terminal term remains primary, while C_j supplies generator-aware auxiliary shaping; λ_F is fixed on a disjoint development split.

Reliability-gated generator routing. Let $D_t = [\bar{d}_{j\leftarrow k,t}]_{j\neq k}$. The sampled current-policy response operator is a trace-normalized ridge map. To estimate whether this finite span generalizes beyond the $K - 1$ rollout candidates, we sample B_{train} additional intent-preserving prompt alternatives at the

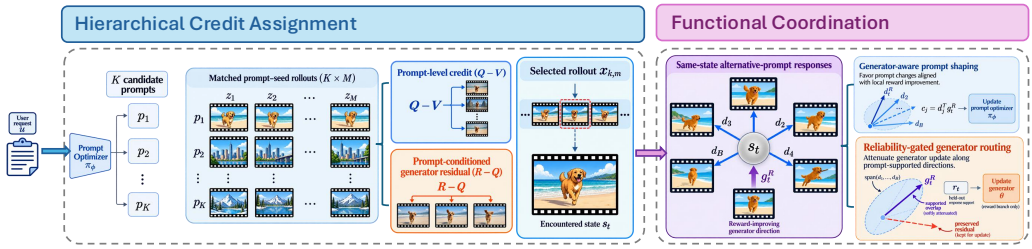


Figure 2: **FLOWCRED training overview.** Hierarchical credit assignment samples K prompt candidates and reuses M matched generator seeds, so cross-prompt contrasts estimate prompt-level credit $Q - V$ while within-prompt residuals estimate $R - Q$. A selected rollout supplies an encountered state s_t for functional coordination. At that fixed state, alternative-prompt responses define a local geometry used for generator-aware prompt shaping and reliability-gated routing of only the generator reward branch.

same state. Each requires only one same-state generator transition forward: no video is decoded and no reward model is called. With response $\widetilde{d}_b$, define

$$
\begin{aligned}
P_{\lambda,t} &= D_t(D_t^\top F_t D_t + \lambda_t I)^{-1} D_t^\top F_t, & \lambda_t &= \epsilon_{\text{ridge}} \frac{\operatorname{tr}(D_t^\top F_t D_t)}{K-1}, \\
r_t &= \operatorname{clip}\left(\frac{1}{B_{\text{train}}} \sum_b \frac{\|P_{\lambda,t}\widetilde{d}_b\|_{F_t}^2}{\|\widetilde{d}_b\|_{F_t}^2 + \epsilon_{\text{num}}}, 0, 1 \right), & \alpha_t &= \alpha_{\max} r_t, \\
g_t^G &= (I - \alpha_t P_{\lambda,t}) g_t^R, & q_t^G &= F_t g_t^G.
\end{aligned}
\tag{8}
$$

Here ϵ_{ridge} is the fixed ridge scale, $\alpha_{\max}$ caps the routing strength, and q_t^G is the routed reward covector backpropagated through the generator. The score r_t measures held-out response support rather than reward confidence. Low-support states continuously recover the base reward step. For stochastic-transition updaters such as Flow-GRPO, g_t^R is a stochastic estimate of the local reward gradient; the response operator and support gate are built from the cached state and detached prompt responses. Consequently the same linear routing applies to the conditional expected reward gradient, rather than implying that one high-dimensional transition sample itself has large prompt-supported energy. Appendix A.4 states this distinction explicitly. The same geometry therefore serves two targets: C_j shapes the upstream prompt action, whereas Eq. 8 regularizes the downstream expected reward update without assuming an exact transfer between optimizers.

3.4 TRAINING PROCEDURE

One transition per decoded trajectory is routed by default, sampled uniformly from the valid stochastic-policy steps used by the base reward loss; all other reward-loss transitions and the entire regularization/reference branch remain unchanged. Same-state quantities are stop-gradient targets from the detached behavior snapshot and are recomputed online each round. Default hyperparameters and backend-specific instantiations are in Appendices D and A.4.

Algorithm 1 One joint post-training round

Require: request u; prompt policy π_ϕ; generator G_θ; K candidates; M matched seeds

Hierarchical credit assignment

1: sample $p_{1:K}$ and decode $x_{k,m} = G_\theta(p_k; z_m)$ with shared $z_{1:M}$
2: evaluate $R_{k,m}$; compute $\widehat{A}_k^{P,R}$ and $\widehat{A}_{k,m}^{G}$

Functional coordination at actual rollout states

3: **for** each selected transition state $s_t^{k,m}$ **do**
4: obtain (q_t^R, g_t^R) from the hierarchy-weighted base reward branch
5: probe $d_{j \leftarrow k,t}$ for $j \neq k$ and accumulate cross-fitted evidence C_j
6: construct $P_{\lambda,t}$; estimate r_t with B_{train} non-decoded prompt probes
7: route $g_t^G \leftarrow (I - \alpha_{\max} r_t P_{\lambda,t}) g_t^R$
8: **end for**
9: update prompt policy with Eq. 7; update generator with the routed reward branch plus unchanged regularization

4 EXPERIMENTS

4.1 PROTOCOL AND CONTROLLED COMPARISONS

Setup. Our primary system uses LTX-2.5 (HaCohen et al., 2026; Lightricks, 2026), a Qwen3.5-9B prompt optimizer, Flow-GRPO, and $K = M = 4$. Controlled variants share initialization, request stream, prompt sampling, matched seeds, reward, decoded-video budget, and optimization schedule; functional coordination adds same-state transition forwards but no extra video decode. The frozen training reward uses public, fixed scorers with weights $0.35/0.30/0.25/0.10$: the Text-Alignment head of VideoReward (Liu et al., 2025b) for request faithfulness, frame-averaged VQAScore (CLIP-FlanT5-XXL) (Lin et al., 2024) for compositional correspondence, the temporal-consistency/text-alignment aggregate of VideoScore-v1.1 (He et al., 2024) for event realization, and the visual-/motion-quality aggregate of VideoReward as a bounded quality stabilizer. All scorers and calibration statistics are frozen before controlled training. Appendices D and D.3 specify checkpoints, frame sampling, calibration, and scorer-family robustness.

Evaluation and comparisons. We report VBench (Huang et al., 2024), VBench-2.0 (Zheng et al., 2025), T2V-CompBench (Sun et al., 2025), and StoryEval (Wang et al., 2025) on the original request u with matched evaluation seeds. The hierarchical-credit-versus-Full comparison and the functional-coordination factorial use three independent training runs; other trainable baselines use two. The controlled chain spans prompt-only adaptation, a raw-request Flow-GRPO generator-only reference, PromptRL, UniGRPO, Matched Global-LOO, hierarchical credit, and Full. For methods originally reported with different backbones or training stacks, we implement the published objective in the common Qwen3.5/LTX-2.5 harness; the table therefore reports controlled common-harness results rather than copied published absolute scores. CAPE-T2V is positioned separately because its public LTX recipe uses sequential two-sided adaptation with additional video-derived caption supervision and a different released LTX backbone; Appendix D.5 gives the precise comparison scope. Uncertainty and data-integrity protocols are summarized in Appendices D.4, G, and J.

4.2 MAIN BENCHMARK PERFORMANCE

Table 1 compares single-side and joint alternatives under one LTX-2.5 harness. Because prompt-side adaptation already closes substantial interface-level headroom, the matched joint chain is the key test of hierarchical credit and functional coordination.

Table 1: **Main LTX-2.5 comparison under a common evaluation harness.** Mean scores use matched evaluation seeds; training-run uncertainty is reported in Appendix J. Best is bold and second-best is underlined.

Method	VBench↑	VBench-2.0↑	T2V-CompBench↑	StoryEval↑
FROZEN / INTERFACE REFERENCES				
Raw prompt	78.14	54.52	46.52	37.88
Official LTX-2.5 prompt enhancer	84.92	61.36	57.48	51.63
SINGLE-SIDE ADAPTATION				
Prompt-A-Video	85.63	62.18	58.87	52.96
VPO	85.44	<u>64.06</u>	59.54	53.61
Prompt-only	<u>86.97</u>	62.84	60.28	54.37
Flow-GRPO	80.83	58.57	52.31	45.26
JOINT POST-TRAINING — MATCHED DECODED-ROLLOUT BUDGET				
PromptRL	86.31	63.02	61.06	55.02
UniGRPO	86.40	63.28	61.32	55.40
Matched Global-LOO	86.49	63.41	61.47	55.58
FLOWCRED (hierarchical credit)	86.91	63.92	<u>62.63</u>	<u>56.71</u>
FLOWCRED (Full)	**87.56**	**64.73**	**64.46**	**58.08**

Prompt-side adaptation already closes substantial interface-level headroom, making the matched joint chain the stricter test. In particular, Prompt-only is already competitive on VBench, whereas the value of joint adaptation is more visible on compositional and multi-event metrics: Full exceeds

Prompt-only by +4.18 T2V-CompBench and +3.71 StoryEval points. Within the matched joint chain, PromptRL and UniGRPO remain below Matched Global-LOO; hierarchical credit then improves over Global-LOO by +0.42/+0.51/+1.16/+1.13 on VBench/VBench-2.0/T2V-CompBench/StoryEval, and Full adds another +0.65/+0.81/+1.83/+1.37. The hierarchical-credit-to-Full increment is positive in all three independent training runs on every headline benchmark. Appendix J reports run-level deltas and intervals, and Table 8 gives the aggregate breakdown.

4.3 Hierarchical-credit diagnostics

The two estimator choices are directly testable at fixed decoded-video budget. Conditioning the generator baseline on the executed prompt removes prompt-level nuisance variation from the downstream advantage, while matched seeds pair prompt candidates under common generator randomness. Table 2 reports held-out diagnostics; the corresponding distributions and protocol are in Appendix E. Together, these diagnostics support the intended separation of prompt-level and prompt-conditioned variation while keeping the decoded-rollout budget fixed.

Table 2: **Hierarchical-credit diagnostics at matched rollout budget.** Lower variance is better; the final row keeps the decoded-video count fixed.

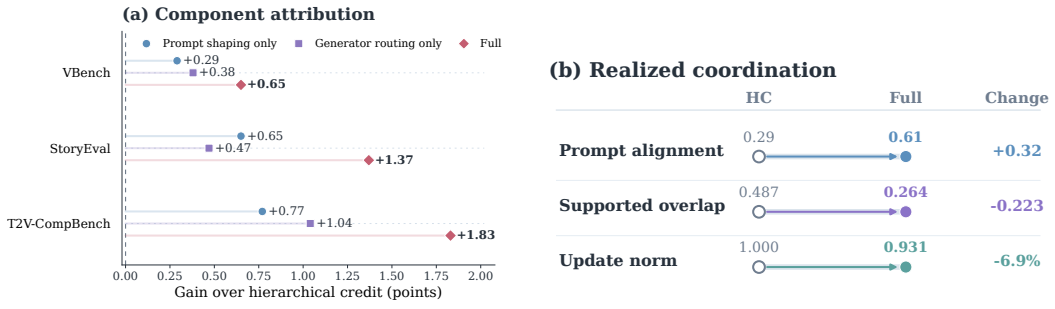
Diagnostic	Reference	Hierarchical design	Result
Generator-advantage variance	Global LOO	Prompt-conditioned LOO	-31%
Prompt-credit variance	Independent seeds	Matched seeds	-27%
T2V-CompBench	Independent seeds	Matched seeds	$62.21 \rightarrow 62.63$

4.4 Functional-coordination attribution and mechanism diagnostics

The two uses of the same-state geometry contribute independently, and their combination yields the strongest result. Table 3 reports the three-run mean (SD) endpoint scores, while Fig. 3 summarizes the complementary endpoint-attribution and realized-update audits. Relative to hierarchical credit, prompt shaping alone improves VBench/StoryEval/T2V-CompBench by +0.29/+0.65/+0.77, generator routing alone by +0.38/+0.47/+1.04, and Full by +0.65/+1.37/+1.83.

Table 3: **Functional-coordination factorial ablation.** Mean (SD) across the same three-run protocol used for the HC–Full comparison.

Prompt shaping	Generator routing	VBench↑	StoryEval↑	T2V-CompBench↑
–	–	86.91 (0.14)	56.71 (0.27)	62.63 (0.32)
✓	–	87.20 (0.15)	57.36 (0.28)	63.40 (0.34)
–	✓	87.29 (0.13)	57.18 (0.26)	63.67 (0.31)
✓	✓	**87.56 (0.12)**	**58.08 (0.24)**	**64.46 (0.29)**



(a) Component attribution. (b) Realized coordination.

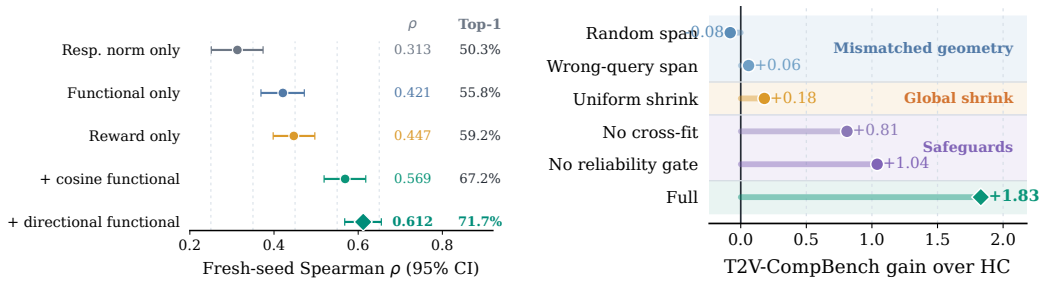
Figure 3: **Functional-coordination endpoint and mechanism evidence. (a)** Relative to hierarchical credit, prompt shaping and generator routing each improve endpoint performance, and their combination is strongest on all three metrics. **(b)** In a matched cached-state audit, Full increases prompt-side alignment, reduces supported overlap, and retains most of the overall generator update norm.

Realized coordination. The factorial result is complemented by a matched one-step audit on cached states. Rather than interpreting the projection energy of an individual stochastic policy-gradient sample, we measure the neural displacement after the optimizer step. Full reduces realized generator supported overlap from 0.487 to 0.264 while retaining 0.931× the hierarchical-credit update norm; on the prompt side, the native-metric alignment cosine rises from 0.29 to 0.61. Shaping-only and routing-only variants isolate the corresponding upstream and downstream changes (Appendix K). Together with the endpoint factorial, these measurements support selective coordination rather than a generic shrinkage explanation, while remaining descriptive checks of trained behavior rather than an exact transfer of causal responsibility.

4.5 Prompt-policy signal validity and routing controls

On 360 held-out requests, reward-only candidate ranking reaches $\rho = 0.447$ / 59.2% Top-1 on eight disjoint fresh seeds, cosine functional evidence reaches $\rho = 0.569$ / 67.2%, and the directional functional score used by FlowCred reaches $\rho = 0.612$ / 71.7%. This provides an out-of-seed check that same-state generator information carries signal beyond terminal reward ranking alone.

Matched-budget controls test simpler routing explanations. Rank-matched random spans and wrong-query spans stay near the hierarchical-credit reference; uniform shrink yields only a small gain, while removing cross-fitting or the reliability gate reduces the Full improvement. Figure 4 summarizes both checks; complete control values are reported in Appendix H.



(a) Fresh-seed predictive utility. (b) Compact matched-budget routing controls.

Figure 4: **Functional signal validity and routing controls.** Directional evidence improves fresh-seed candidate ranking beyond reward-only scoring, while compact routing controls test mismatched geometry and safeguard removal.

4.6 Generalization and external validation

Transfer evidence. We transfer the functional-coordination increment to HunyuanVideo-1.5 and to RVM and Diffusion-DRF generator updates. Table 4 reports matched Full-minus-hierarchical-credit deltas rather than heterogeneous absolute rankings.

Table 4: **Transfer of functional coordination.** Full minus matched hierarchical credit under each backbone/updater.

Setting	ΔVBench	ΔT2V-CompBench	ΔStoryEval
LTX-2.5 / Flow-GRPO	+0.65	+1.83	+1.37
HunyuanVideo-1.5 / Flow-GRPO	+0.57	+1.22	+1.09
LTX-2.5 / RVM	+0.52	+1.29	+1.15
LTX-2.5 / Diffusion-DRF	+0.41	+0.97	+0.79

The increment remains positive in every tested backbone/updater setting, although its magnitude varies, indicating that functional coordination is not tied to a particular generator backbone or reward-update rule.

Full keeps the decoded-video and reward-model-call budget identical to hierarchical credit; its additional generator work consists of non-decoded same-state transition probes, counted explicitly

in Appendix I. Frozen-module swaps attribute measurable gains to both sides: substituting the Full generator into the hierarchical-credit pair adds +0.90 T2V-CompBench and substituting the Full prompt policy adds +0.72. In a 240-request blinded evaluation, Full is preferred over hierarchical credit on 134 versus 48 requests (58 ties; 73.6% of non-ties) and over PromptRL on 142 versus 46 (52 ties; 75.5% of non-ties). An independent VisionReward evaluator, never used in training or model selection, also prefers Full to hierarchical credit on 63.3% of 300 held-out matched pairs. Detailed annotation and replication protocols are in Appendices J and L.

Taken together, these checks test complementary explanations of the primary gain. Transfer across a second backbone and alternative updater families probes dependence on the default generator stack; frozen-module swaps verify contributions from both trained modules; and blinded human judgments plus the held-out VisionReward evaluator test whether the improvement is confined to the training reward.

External two-sided positioning. PromptRL and UniGRPO use their published joint-training objectives in the common LTX-2.5 harness. CAPE-T2V is reported only as an external two-sided reference because its public pipeline uses a different released LTX backbone and auxiliary video-derived captions. Appendix D.5 details the comparison scope; Appendix J reports rollout-budget and paired-run checks.

4.7 Qualitative coordination cases

Figure 5 shows two representative cases with complementary demands. In the structured repair scene, Base and VPO capture the broad setting but do not make the requested spacecraft-repair interaction equally explicit, whereas FlowCred preserves both the scene structure and the target interaction. In the directional-motion case, the supplementary full video shows clearer execution of the requested right-to-left motion for FlowCred, while the frame strip highlights its subject consistency; the alternatives show weaker directional execution or greater subject ambiguity. The corresponding full videos are included in the supplementary material; Appendix L provides four additional cases spanning counting, interaction, spatial relation, and synchronized appearance change.

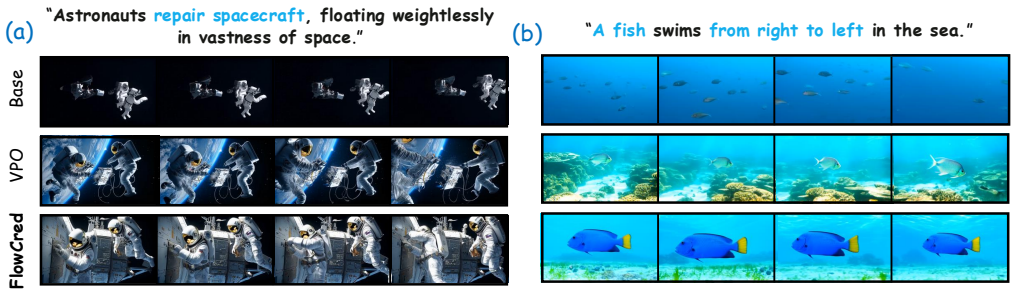


Figure 5: **Representative qualitative comparison.** Base, VPO, and FlowCred on structured scene realization and directional motion. Corresponding full videos are provided in the supplementary material.

5 Conclusion

We presented FlowCred, a framework for jointly post-training a prompt optimizer and a text-to-video generator without treating a shared terminal reward as undifferentiated supervision. Hierarchical credit assignment uses matched prompt–seed rollouts to separate prompt choice from prompt-fixed stochasticity, while functional coordination uses same-state response geometry for generator-aware prompt shaping and reliability-gated routing of the generator reward branch. Controlled comparisons show that the gains cannot be explained by stronger prompts, additional decoded rollouts, or generic update shrinkage alone, and evidence from repeated runs, mechanism analyses, human evaluation, and transfer studies is consistent with the intended division of labor between prompt-side conditioning and residual generator execution. More broadly, FlowCred suggests that joint post-training should distinguish *where outcome variation arises* from *where trainable modules can realize overlapping functional changes*: sharing a reward is not sufficient by itself; credit assignment and update coordination are separate design problems in cascaded generative systems.

AI USE STATEMENT

Generative AI tools were used to assist with conceptual and experimental design, literature search, code development, and manuscript drafting and editing. All AI-assisted material was reviewed by the authors, who are responsible for the final submission.

REPRODUCIBILITY STATEMENT

Section 3 specifies the FLOWCRED formulation and training procedure; Appendices A–L provide derivations, implementation details, reward/calibration and baseline protocols, evaluation, compute, robustness, human evaluation, and qualitative results. The supplementary material includes an anonymous reference implementation and the full videos for the two main-text qualitative cases.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. BeautifulPrompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1–11. Association for Computational Linguistics, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. VPO: Aligning text-to-video generation models with prompt optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15636–15645, 2025.

Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. Video-as-answer: Predict and generate next video event with Joint-GRPO. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38915–38925, 2026.

Jaemoo Choi, Wei Guo, Yuchen Zhu, Arash Vahdat, Molei Tao, Julius Berner, and Yongxin Chen. Scaling reinforcement learning for diffusion models via velocity matching. *arXiv preprint arXiv:2608.23664*, 2026.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024.

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. In *Advances in Neural Information Processing Systems*, 2023.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2974–2982, 2018. doi: 10.1609/aaai.v32i1.11794.

Yuanshen Guan, Zipeng Feng, Chengru Song, Zhiwei Xiong, and Peiqin Sun. Latent reward registers for diffusion preference alignment. *arXiv preprint arXiv:2608.03929*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate

fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123. Association for Computational Linguistics, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 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.

Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, Weilin Huang, and Ping Luo. Prompt-A-Video: Prompt your video diffusion model via preference-aligned llm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18725–18735, 2025.

Yizhuo Jia, Jingyun Hua, and Yuanxing Zhang. CAPE-T2V: Captioner-anchored prompt enhancement toward two-sided conditioning alignment in text-to-video generation. *arXiv preprint arXiv:2608.03046*, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Lightricks. LTX-2.5: Video, audio & world simulation. Hugging Face model card, 2026. URL `https://huggingface.co/Lightricks/LTX-2.5`. Accessed September 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024*, pp. 366–384, 2024. doi: 10.1007/978-3-031-72673-6_20.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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. In *Advances in Neural Information Processing Systems*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025b.

Jie Liu, Zilyu Ye, Linxiao Yuan, Shenhan Zhu, Yu Gao, Jie Wu, Kunchang Li, Xionghui Wang, Xiaonan Nie, Weilin Huang, and Wanli Ouyang. UniGRPO: Unified policy optimization for reasoning-driven visual generation. *arXiv preprint arXiv:2603.23500*, 2026.

Yifu Luo, Haoyuan Sun, Xinhao Hu, Penghui Du, Keyu Fan, Bo Li, Sinan Du, Xu Wan, Zhiyu Chen, Bo Xia, Yongzhe Chang, Changqian Yu, Kun Gai, Tiantian Zhang, and Xueqian Wang. Principled rl for flow matching emerges from the chunk-level policy optimization. In *International Conference on Machine Learning*, 2026. arXiv:2510.21583.

Andrew Y. Ng and Michael I. Jordan. Pegasus: A policy search method for large mdps and pomdps. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pp. 406–415, 2000.

Daisuke Oba, Hiroki Furuta, and Naoaki Okazaki. Diffusion-state policy optimization for masked diffusion language models. *arXiv preprint arXiv:2602.06462*, 2026.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024.

Yujiang Pu and Yu Kong. Noiseasier: Test-time noise optimization for text-to-video generation. In *European Conference on Computer Vision*, 2026. arXiv:2608.30194.

Qwen Team. Qwen3.5-9B. Hugging Face model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-9B`. Accessed September 2026.

Yash Savani, Branislav Kveton, Yuchen Liu, Yilin Wang, Jing Shi, Subhojyoti Mukherjee, Nikos Vlassis, and Krishna Kumar Singh. Stepwise credit assignment for grpo on flow-matching models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42007–42017, 2026.

Jinjie Shen, Wei Deng, Xian Hu, Daiguo Zhou, and Jian Luan. STAR: Spatiotemporal adaptive reward allocation for text-to-image rl post-training. *arXiv preprint arXiv:2606.17979*, 2026.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 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 *Proceedings of the IEEE/CVF 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.

Fu-Yun Wang, Han Zhang, Michael Gharbi, Hongsheng Li, and Taesung Park. PromptRL: Prompt matters in rl for flow-based image generation. *arXiv preprint arXiv:2602.01382*, 2026a.

Yifan Wang, Yanyu Li, Gordon Guocheng Qian, Sergey Tulyakov, Yun Fu, and Anil Kag. Diffusion-DRF: Free, rich, and differentiable reward for video diffusion fine-tuning. *arXiv preprint arXiv:2601.04153*, 2026b.

Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13629–13638, 2025.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, et al. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

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

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models. In *International Conference on Machine Learning*, 2026. arXiv:2509.25050.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Xi Ye, Wenjia Yang, Yangyang Xu, Xiaoyang Liu, Duo Su, Mengfei Xia, and Jun Zhu. SHIFT: Motion alignment in video diffusion models with adversarial hybrid fine-tuning. In *European Conference on Computer Vision*, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Jingyuan Zhu, Biaolong Chen, Le Zhang, Aixi Zhang, Hao Jiang, and Pipei Huang. Diffusion-APO: Trajectory-aware direct preference alignment for video diffusion transformers. *arXiv preprint arXiv:2605.07503*, 2026.

A THEORETICAL FOUNDATIONS

This appendix collects the mathematical statements used in the main paper. The reward hierarchy is exact for any per-trajectory scalar terminal score and yields valid score-function baselines. In the final protocol, reward-component and prompt-term affine scales are estimated once on disjoint training-only calibration data and then frozen, so current minibatch composition does not enter the score used by the hierarchy (Appendix D.3). Same-state prompt intervention provides a first-order directional derivative of the generator reward surrogate. Minimum-overlap routing is a local function-space regularizer; its non-reversal statement is made in the backend's local output coordinates before neural pullback.

A.1 HIERARCHY-AWARE STATISTICAL CREDIT

Recall

$$J(\phi,\theta)=\mathbb{E}_{u\sim\mathcal{D}}\mathbb{E}_{p\sim\pi_\phi(\cdot|u)}\mathbb{E}_{\mathcal{T}\sim\pi_\theta(\cdot|p)}[R(u,x)], \tag{9}$$

with

$$Q_\theta(u,p)=\mathbb{E}_{\mathcal{T}\sim\pi_\theta(\cdot|p)}[R(u,x)],\qquad V_{\phi,\theta}(u)=\mathbb{E}_{p\sim\pi_\phi(\cdot|u)}Q_\theta(u,p). \tag{10}$$

Proposition 1 (Hierarchy-aware score-function baselines)**.** *For the factorized policy $\pi_\phi(p|u)\pi_\theta(\mathcal{T}|p)$,*

$$\nabla_\phi J=\mathbb{E}[(Q_\theta(u,p)-V_{\phi,\theta}(u))\nabla_\phi\log\pi_\phi(p|u)], \tag{11}$$

$$\nabla_\theta J=\mathbb{E}[(R(u,x)-\mathrm{sg}[Q_\theta(u,p)])\nabla_\theta\log\pi_\theta(\mathcal{T}|p)]. \tag{12}$$

Proof. For the prompt policy, apply the score-function identity and subtract $V(u)$, which does not depend on the sampled prompt action. For the generator, any prompt-dependent quantity that is fixed with respect to the sampled downstream trajectory is a valid score-function baseline; choosing $Q(u,p)$ gives the second expression. The stop-gradient indicates that the statement concerns a policy-gradient baseline rather than differentiation through the value estimate. □

The pointwise identity $R-V=(Q-V)+(R-Q)$ follows immediately. It is a statistical decomposition of reward variation. Functional coordination in Section 3.3 then shapes the joint objective using same-state response evidence, so the two levels have distinct mathematical status.

A.2 RELATION TO THE MATCHED GLOBAL-LOO CONTROL

The matched control uses the same sampled actions and matched rollout grid as hierarchical credit, but both modules receive a global leave-one-out advantage

$$A^{\mathrm{GLOO}}_{k,m}=R_{k,m}-\frac{1}{KM-1}\sum_{(j,n)\neq(k,m)}R_{j,n}. \tag{13}$$

For a fixed prompt action p_k, averaging over its M trajectories yields a positive affine rescaling of $Q_k-\bar{Q}$ and therefore preserves the coarse prompt ordering. The hierarchical-credit prompt-policy estimator in Eq. 4 is likewise a positive rescaling of $Q_k-\bar{Q}$. We consequently do not attribute the hierarchical-credit gain to a new prompt ordering. The controlled change is the downstream generator baseline: Global-LOO removes only the current trajectory, whereas hierarchical credit removes prompt-level variation as well by comparing the trajectory against other seeds under the same executed prompt. The global-LOO construction also eliminates the self-inclusion confound that would otherwise accompany a global mean baseline.

A.3 VARIANCE REDUCTION FROM MATCHED SEEDS

For candidates p_k,p_j, write $R_k(z)$ and $R_j(z)$ for rewards obtained with common seed z. Define

$$\widehat{\Delta}^{\mathrm{match}}_{kj}=\frac{1}{M}\sum_{m=1}^{M}[R_k(z_m)-R_j(z_m)]. \tag{14}$$

For an independently seeded comparison, use z_m for k and independent z'_m for j. Let $\sigma_k^2=\mathrm{Var}(R_k)$, $\sigma_j^2=\mathrm{Var}(R_j)$, and $\sigma_{kj}=\mathrm{Cov}(R_k(z),R_j(z))$.

Proposition 2 (Matched-seed variance)**.** *Assume independent seed pairs across m and finite second moments. Then*

$$\mathrm{Var}(\widehat{\Delta}_{kj}^{\mathrm{match}}) = \frac{\sigma_k^2 + \sigma_j^2 - 2\sigma_{kj}}{M}, \tag{15}$$

$$\mathrm{Var}(\widehat{\Delta}_{kj}^{\mathrm{ind}}) = \frac{\sigma_k^2 + \sigma_j^2}{M}. \tag{16}$$

Hence matched seeds reduce variance by $2\sigma_{kj}/M$ whenever $\sigma_{kj} \geq 0$.

Proof. Apply the variance identity to $R_k(z) - R_j(z)$ and use independence across m. The independent estimator has zero cross-covariance by construction. □

The prompt-policy estimator in Eq. 4 averages these seed-wise contrasts over the other $K - 1$ candidates, so common randomness enters the actual training estimator rather than only an auxiliary analysis.

A.4 Reward-only transition geometry

For a Gaussian transition $\pi(a|\mu) = \mathcal{N}(\mu, \Sigma)$ with covariance independent of μ,

$$\nabla_\mu \log \pi(a|\mu) = \Sigma^{-1}(a - \mu), \qquad F = \mathbb{E}[\nabla_\mu \log \pi \nabla_\mu \log \pi^\top] = \Sigma^{-1}. \tag{17}$$

If an unclipped reward surrogate has local coefficient $\widehat{A}$, then $q^R = \widehat{A}\Sigma^{-1}(a - \mu)$ and $g^R = F^{-1}q^R = \widehat{A}(a - \mu)$. This provides intuition for the Fisher coordinate used in the main text. The implementation differentiates the actual reward branch of Flow-GRPO after its ratio and clipping operations. Reference/KL terms stay inside J_{reg} and never enter q^R, preventing reward credit from being contaminated by a return-to-reference direction.

Stochastic sample versus expected reward direction. For Flow-GRPO, the single-transition vector above is a stochastic score-function sample, so its high-dimensional direction need not itself align strongly with a low-rank prompt-response span. Let P_t and α_t be constructed from the cached pre-action state and detached prompt responses, hence fixed with respect to the current transition draw conditional on that state. Linearity gives

$$\mathbb{E}[(I - \alpha_t P_t) g_t^R \mid s_t] = (I - \alpha_t P_t)\, \mathbb{E}[g_t^R \mid s_t]. \tag{18}$$

Thus routing is interpreted as a regularizer on the conditional expected reward update. We do not use the projection energy of one stochastic transition sample as a semantic-overlap statistic; the mechanism audits below instead use realized neural displacements after the optimizer step.

A.5 Directional prompt-policy evidence

Fix a behavior state s_t, executed action p_k, and differentiable local reward surrogate $J_{\mathrm{reward,t}}(\mu)$. Let $\mu_k = \mu(s_t, p_k)$, $q_t^R = \nabla_\mu J_{\mathrm{reward,t}}(\mu_k)$, and $d = \mu(s_t, p_j) - \mu_k$. Taylor's theorem gives

$$J_{\mathrm{reward,t}}(\mu_k + \varepsilon d) = J_{\mathrm{reward,t}}(\mu_k) + \varepsilon (q_t^R)^\top d + \frac{\varepsilon^2}{2} d^\top H(\xi) d. \tag{19}$$

Therefore $c = d^\top q_t^R$ is exactly the directional derivative at $\varepsilon = 0$. It characterizes the local reward change along an observed same-state conditioning response; fresh-seed evaluation measures how far that local signal transfers to full trajectories. The held-out-response and fresh-seed experiments in Appendix F test whether this local quantity carries predictive information beyond the construction states.

A.6 Minimum-overlap routing in Fisher coordinates

For positive-definite F, define $\langle a, b \rangle_F = a^\top F b$ and $\|a\|_F^2 = a^\top F a$. Let P be the F-orthogonal projector onto the response span,

$$P = D(D^\top F D)^\dagger D^\top F. \tag{20}$$

Then $P^2 = P$ and $P^\top F = FP$.

Proposition 3 (Minimum-overlap solution). *For $\eta \geq 0$, the local problem*

$$\min_{\delta} \ \frac{1}{2}\|\delta - g^R\|_F^2 + \frac{\eta}{2}\|P\delta\|_F^2 \tag{21}$$

has solution

$$\delta^* = (I - \alpha P)g^R, \qquad \alpha = \frac{\eta}{1+\eta}, \tag{22}$$

and

$$(q^R)^\top \delta^* = \|(I-P)g^R\|_F^2 + (1-\alpha)\|Pg^R\|_F^2 \geq (1-\alpha)\|g^R\|_F^2 \geq 0. \tag{23}$$

Proof. The first-order condition is $F(\delta - g^R) + \eta P^\top F P\delta = 0$. Using $P^\top F = FP$ and $P^2 = P$ yields $(I + \eta P)\delta = g^R$. Its inverse is $I - \eta(1+\eta)^{-1}P$. The final expression follows from the F-orthogonal decomposition $g^R = Pg^R + (I-P)g^R$. □

The statement is made in an unconstrained transition-mean coordinate. The neural update in Eq. 8 is a pullback through $J_{\mu,\theta}$; an arbitrary mean-space displacement need not be exactly realizable by the parameterization. We therefore use the proposition to motivate the local routing rule and test the realized parameter update directly through the audit in Appendix F.1.

A.7 Ridge stabilization and contraction

The implemented operator in Eq. 8 uses ridge regularization. Let $B = F^{1/2}D$. In whitened coordinates,

$$\widetilde{P}_\lambda = B(B^\top B + \lambda_t I)^{-1}B^\top, \qquad \lambda_t = \epsilon_{\text{ridge}}\,\frac{\operatorname{tr}(D_t^\top F_t D_t)}{K-1}, \quad \epsilon_{\text{ridge}} = 10^{-3}. \tag{24}$$

If $B = U\,\operatorname{diag}(s_i)V^\top$, then

$$\widetilde{P}_\lambda = U\,\operatorname{diag}\left(\frac{s_i^2}{s_i^2 + \lambda_t}\right)U^\top. \tag{25}$$

Hence $\widetilde{P}_\lambda$ is symmetric positive semidefinite with eigenvalues in $[0,1)$. For $0 \leq \alpha \leq 1$ and $\widetilde{g} = F^{1/2}g^R$,

$$(q^R)^\top(I - \alpha P_\lambda)g^R = \widetilde{g}^\top(I - \alpha\widetilde{P}_\lambda)\widetilde{g} \geq (1-\alpha)\|\widetilde{g}\|_2^2 \geq 0. \tag{26}$$

This covers the stabilized local operator used in training without calling it an exact projector.

A.8 Reference action and finite response span

The conditioning reference is the executed prompt action p_k because it is the upstream action that produced s_t. The intervention of interest is therefore $p_k \to p_j$. Using the raw request u as the reference mixes candidate-specific variation with the shared raw-to-enhanced displacement and can overstate directions common to all enhanced prompts; we keep that choice as an ablation.

With $K = 4$, the main response matrix has at most three columns, so the geometry is a finite, policy-dependent response span. It measures overlap with alternatives actually proposed by the current prompt policy. Rank-matched random spans, wrong-query spans, held-out response prediction, and generator swaps are designed around that finite claim.

B Reliability, Trust Regions, and Detachment

For selected states, the prompt policy samples $B_{\text{train}} = 6$ additional intent-preserving textual alternatives $\widetilde{p}_b$ that are not rendered into videos. The main span is always constructed from the $K-1$ real rollout candidates; these non-decoded probes require only one same-state generator transition forward and no reward-model call. For $\widetilde{d}_b = \mu(s_t, \widetilde{p}_b) - \mu(s_t, p_k)$, define

$$e_{b,t} = \frac{\|P_{\lambda,t}\widetilde{d}_b\|_{F_t}^2}{\|\widetilde{d}_b\|_{F_t}^2 + \epsilon_{\text{num}}}, \qquad r_t = \operatorname{clip}\left(\frac{1}{B_{\text{train}}}\sum_{b=1}^{B_{\text{train}}} e_{b,t}, 0, 1\right), \qquad \alpha_t = \alpha_{\max} r_t, \tag{27}$$

with $\epsilon_{\text{num}} = 10^{-12}$. The score measures support/coverage of the sampled response span, not correctness of the reward direction. The default uses $\alpha_{\max} = 0.75$. Evaluation-time mechanism analysis independently samples $B_{\text{eval}} = 12$ probes on held-out requests; these probes never participate in the training gate.

For a response d, the trust operator is

$$\text{TrustClip}_{F_t}(d) = d \, \min\left(1, \sqrt{\frac{\tau_{\text{batch}}}{d^\top F_t d + \epsilon_{\text{num}}}}\right), \tag{28}$$

where τ_{batch} is the detached 95th percentile of native-metric response energies $d^\top F_t d$ in the current training batch. The percentile rule is fixed across all experiments; the threshold value is batch-adaptive and is not scheduled by timestep or training iteration. Thus in-region responses are unchanged and only unusually large conditioning shifts are rescaled. Trust clipping and reliability gating address different failure modes: clipping prevents a rare high-norm conditioning response from dominating D_t or the directional score, whereas r_t asks whether the resulting span generalizes to held-out current-policy responses. We apply three safeguards before a response enters D_t: (i) prompt-level intent filtering rejects alternatives that alter the requested event or introduce unsupported entities; (ii) Eq. 28 limits response magnitude; and (iii) near-collinear columns are stabilized by the trace-normalized ridge term rather than explicitly inverted. The percentile rule is batch-adaptive and is used as a robustifier rather than a separately tuned reward weight; fixed-α and binary-gate controls isolate the subsequent routing logic.

The behavior generator snapshot θ^- is detached when computing alternative responses. Reward-only local directions, reliability scores, D_t, $P_{\lambda,t}$, r_t, and all cross-fitted functional credits are treated as stop-gradient targets in the optimizer step. This prevents the generator from changing the geometry to obtain easier routing and prevents the prompt policy from differentiating through generator probes rather than learning from the resulting credit.

C Training Algorithm

Algorithm 2 One joint post-training round

Require: request u; prompt policy π_ϕ; generator G_θ; K candidates; M matched seeds

Hierarchical credit assignment

sample $p_{1:K}$ and decode $x_{k,m} = G_\theta(p_k; z_m)$ with shared $z_{1:M}$

evaluate $R_{k,m}$ against the original request u; compute $\widehat{A}^{P,R}_k$ and $\widehat{A}^{G}_{k,m}$

Functional coordination at actual rollout states

for each selected actual transition state $s_t^{k,m}$ **do**

  obtain reward-only local direction (q_t^R, g_t^R) from the base generator updater

  probe $d_{j\leftarrow k,t}$ for $j \neq k$; cross-fit $d_{j\leftarrow k,t}^\top q_t^R$ into C_j

  construct $P_{\lambda,t}$; estimate support r_t from B_{train} independent non-decoded prompt probes

  route $g_t^G \leftarrow (I - \alpha_{\max} r_t P_{\lambda,t}) g_t^R$

end for

update prompt policy with $\text{Cal}_R(\widehat{A}^{P,R}) + \lambda_F \, \text{Cal}_C(C)$

update generator with $F_t g_t^G$ plus the unchanged base regularizer

D Implementation and Evaluation Protocol

D.1 Backbone, prompt optimizer, and training pool

The primary generator is the public LTX-2.5 trainable DiT checkpoint from the official `Lightricks/LTX-2.5` release (HaCohen et al., 2026; Lightricks, 2026), specifically `diffusion_models/ltx-2.5-22b-dev-transformer-bf16.safetensors`, with its native flow parameterization and scheduler. The prompt optimizer is the instruction-tuned `Qwen/Qwen3.5-9B` checkpoint (Qwen Team, 2026), used as a text-generation policy; we do not use the separate `Qwen3.5-9B-Base` checkpoint. The main configuration uses full-parameter prompt-policy updates. HunyuanVideo-1.5 (Wu et al., 2025) provides backbone transfer, while

RVM (Choi et al., 2026) provides an alternative generator reward updater under the same LTX evaluation harness.

The training pool contains 8,000 T2V requests. We build it from 4,800 programmatically instantiated compositional templates over entities, attributes, relations, counts, and actions; 2,000 independently generated motion/interaction requests; and 1,200 manually curated multi-event requests, then stratify the mixture across object/attribute binding, spatial relation, motion binding, action binding, interaction, numeracy, and longer event descriptions. Before training, requests are deduplicated against every benchmark split by normalized exact match, 5-gram MinHash (Jaccard threshold 0.80), and sentence-embedding cosine (threshold 0.92); no retained training request crosses any of these thresholds. Validation prompts are disjoint at the request level and are never used for benchmark reporting. Prompt candidates are constrained to preserve the user request, avoid unsupported entities, and remain within the generator context budget. Candidate validity is checked by a frozen `Qwen/Qwen3.5-9B` copy that is separate from the trainable prompt-policy weights. The checker receives the original request and one candidate and deterministically returns PASS/FAIL under a fixed rubric: a candidate fails if it changes required entities, counts, relations, actions, or event ordering, or introduces unsupported entities; descriptive camera, lighting, and appearance details are permitted only when they do not alter those constraints. Greedy decoding is used, and the same checker and rubric are applied to all prompt-side baselines and to additional support probes. Rejected candidates are resampled before video generation. For the theory above, π_ϕ denotes the effective distribution over accepted prompt actions induced by this fixed filtering rule. The exactness claim concerns the hierarchical reward decomposition and its score-function baseline identity; the practical prompt-policy update uses the clipped group-relative surrogate in Table 5. The same accepted-action sampling rule is used for all prompt-side methods.

D.2 DEFAULT OPTIMIZATION CONFIGURATION

Table 5: Default LTX-2.5 configuration used by the primary matched comparison. The rows labeled as central controls are held fixed across the principal ablations.

Component	Setting	Role in the comparison
Prompt candidates	$K = 4$	four current-policy actions per request
Matched seeds	$M = 4$	same initial latent and stochastic transition stream per seed across candidates
Decoded rollouts	16/request	held fixed across the central chain
Generator backend	Flow-GRPO	base reward optimizer
Generator trainable set	fixed across generator-training rows	Full changes routing, not which generator parameters are trainable
Joint training rounds	500	global batch 2; 1,000 request draws from the shared 8k pool
Global request batch	2	two requests per joint rollout/update round
Generator optimizer	AdamW, 5×10^{-6}	weight decay 0.01; gradient clip 1.0
Prompt-policy optimizer	AdamW, 1×10^{-6}	weight decay 0.01; gradient clip 1.0
Prompt sampling	$T = 0.8$, top-$p = 0.95$	max 256 new tokens; $K = 4$ accepted candidates
Prompt-policy clip / KL	0.2 / 0.01	group-relative clipped objective with frozen reference
Flow-GRPO clip / KL	10^{-3} / 0.004	one inner epoch; reward and reference branches kept separate
Mixed precision	bfloat16	optimizer states in FP32
Routed states	1/trajectory	uniformly sampled from valid reward-loss transitions
Training support probes	$B_{\text{train}} = 6$	reliability only; no video decode
Candidate scaling	frozen calibration affine scale	Section 3.3; no current-batch statistics
Functional-credit weight	$\lambda_F = 0.35$	combines reward and local prompt credit
Max routing strength	$\alpha_{\max} = 0.75$	upper bound after reliability gating
Trust clip	95th percentile	detached native-metric response energy $d^\top F_t d$; fixed percentile
Ridge coefficient	$\epsilon_{\text{ridge}} = 10^{-3}$	$\lambda_t = \epsilon_{\text{ridge}} \operatorname{tr}(D_t^\top F_t D_t)/(K-1)$; no schedule
Prompt-policy update	full parameter	Qwen3.5-9B policy
Evaluation probes	$B_{\text{eval}} = 12$	mechanism analysis only; independent of training gate
Benchmark seeds	3/request	matched across methods and averaged per request
Central training runs	HC/Full: 3; others: 2	run is the outer replication unit

If the response Gram trace is numerically negligible, we set the response operator and routing strength to zero rather than inverting an ill-conditioned matrix; the clip on r_t is retained only as a numerical guard.

Qualitative outputs use the native 5-second, 24-fps, 832×480 generation format. We apply no spatial upscaling or frame interpolation to the videos or to the frames shown in the paper.

D.3 REWARD AND OPTIMIZATION DETAILS

Each decoded video is scored against the original request by a fixed public reward stack. We use: (i) the Text-Alignment output of `KlingTeam/VideoReward` (Liu et al., 2025b) for request faithfulness r_{faith}; (ii) VQAScore with `zhiqiulin/clip-flant5-xxl` (Lin et al., 2024), averaged over eight uniformly spaced frames, for compositional correspondence r_{comp}; (iii) `TIGER-Lab/VideoScore-v1.1` (He et al., 2024), using the mean of its temporal-consistency and text-to-video-alignment outputs over the model's native frame sampling, for event realization r_{event}; and (iv) the mean of VideoReward visual-quality and motion-quality outputs as the bounded quality stabilizer r_{qual}. All model weights, prompts, frame-sampling rules, and preprocessing are frozen before controlled training.

For each component, location and scale are estimated once on a disjoint calibration pool of 2,048 training-only videos and then held fixed throughout all controlled runs:

$$\widetilde{r}_c = \mathrm{clip}\left(\frac{r_c - \mu_c^{\mathrm{cal}}}{\sigma_c^{\mathrm{cal}} + 10^{-6}}, -3, 3\right), \qquad R^{\mathrm{opt}} = 0.35\widetilde{r}_{\mathrm{faith}} + 0.30\widetilde{r}_{\mathrm{comp}} + 0.25\widetilde{r}_{\mathrm{event}} + 0.10\widetilde{r}_{\mathrm{qual}}. \tag{29}$$

Table 6: Frozen reward-stack configuration and calibration statistics. The calibration pool is disjoint from all controlled training/evaluation requests.

Component	Public checkpoint / signal	Video input	Weight	μ_c^{cal}	σ_c^{cal}
r_{faith}	`KlingTeam/VideoReward`, Text Alignment	native sampler	0.35	0.42	0.17
r_{comp}	`zhiqiulin/clip-flant5-xxl`, VQAScore	8-frame mean	0.30	0.63	0.15
r_{event}	`TIGER-Lab/VideoScore-v1.1`, TC+T2V align	native sampler	0.25	2.71	0.39
r_{qual}	`KlingTeam/VideoReward`, VQ+MQ mean	native sampler	0.10	0.56	0.16

Reported VBench, VBench-2.0, T2V-CompBench, and StoryEval evaluators never enter R^{opt} or the calibration pool. Because $(\mu_c^{\mathrm{cal}}, \sigma_c^{\mathrm{cal}})$ are independent of the current sampled prompts and trajectories, $R^{\mathrm{opt}}(u, x)$ remains a per-trajectory scalar score and Proposition 1 applies directly to the transformed objective. The same principle is used for the frozen calibration transform in Section 3.3: reward-credit and functional-credit scales are calibrated once and frozen rather than recomputed from current candidates.

Scorer-family independence. We additionally evaluate the final hierarchical-credit and Full checkpoints with frozen VisionReward (Xu et al., 2026), which is not used anywhere in training, calibration, model selection, or the four headline benchmarks. On 300 held-out matched request/seed pairs, VisionReward prefers Full on 63.3% of pairs and its normalized scalar score increases by 0.18 standard deviations on average. This scorer-family check is consistent with the blinded human preference study in Appendix L and complements the reward-axis interventions below.

For Flow-GRPO (Liu et al., 2025a), Eq. 4 supplies the hierarchy-aware trajectory advantage used by the reward branch. We use the fixed LTX configuration in Table 5: ratio clip 10^{-3}, KL coefficient 0.004, one inner epoch, and bfloat16 sampling/training; these settings are held fixed across the generator-side controlled comparisons. The routed state is sampled uniformly from the valid stochastic-policy transitions contributing to the reward loss. Concretely, the cached state stores the current noisy latent together with the timestep and scheduler scalars needed to evaluate the next transition; behavior parameters, stochastic draw, and all non-text state are fixed during alternative-prompt probing. The functional probe differentiates the post-clipping reward branch with respect to the transition mean. The reference/KL branch is never routed. Prompt-policy GRPO uses Eq. 7, with terminal and functional components calibrated separately by the frozen statistics defined in Section 3.3 before weighted combination.

Instantiation with RVM. RVM (Choi et al., 2026) uses reward-weighted velocity matching together with an optional anchor toward a reference velocity field. In the reported RVM transfer, functional coordination is instantiated in this backend's native velocity-output coordinates. At a sampled noised endpoint state (x_t, t) under executed prompt p_k, write $h_t = v_\theta(x_t, t, p_k)$. We obtain the local reward-driven direction from the reward-weighted velocity branch, $q_t^R = -\nabla_{h_t}\mathcal{L}_{\mathrm{RVM,reward}}$, absorbing the backend's scalar timestep weight into q_t^R and using the Euclidean metric in h_t coordinates. Alternative-prompt responses use the identical noised state and time under the detached behavior snapshot,

$$d^{\mathrm{RVM}}_{j\leftarrow k,t} = v_{\theta^-}(x_t, t, p_j) - v_{\theta^-}(x_t, t, p_k). \tag{30}$$

The same trust clipping, ridge response operator, and held-out support gate are then applied in velocity-output space. Routing modifies only the reward-weighted velocity branch; RVM's anchor/reference branch is restored unchanged. Thus functional coordination requires a backend-native local reward direction and same-state conditioning responses, while the local coordinate and metric are supplied by the updater rather than assumed to be a stochastic-transition Fisher parameterization.

Instantiation with differentiable reward flow. We additionally instantiate functional coordination on a Diffusion-DRF differentiable reward-flow backend (Wang et al., 2026b). Following the public

DRF setup, the critic is frozen `Qwen2.5-VL-7B`; ten uniformly sampled video frames are used for VLM feedback, and gradients are truncated to the final three denoising steps. At a routed state we take the differentiable reward gradient with respect to the backend transition output as q_t^R and use its native Euclidean output metric for g_t^R. Alternative-prompt responses are evaluated at the identical state under the detached generator snapshot. The same trust clip, trace-normalized ridge operator, and held-out support gate are then applied before backpropagating the routed reward direction. Critic weights, frame sampling, the three-step backpropagation window, gradient checkpointing, and non-reward stabilization terms are held fixed between hierarchy-only and Full. This experiment tests whether cross-module routing remains useful when the generator already receives dense differentiable feedback rather than a scalar policy-gradient advantage.

Backend applicability. Functional coordination requires two backend-local objects: (i) an output coordinate in which changing text conditioning at a fixed state yields a response vector, and (ii) a local improvement covector q_t^R supplied by the generator updater. Flow-GRPO, RVM, and differentiable reward flow provide both directly. A preference objective can be compatible if its implementation exposes a differentiable state-local pairwise loss, in which case the loss gradient can play the role of q_t^R; a purely trajectory-level preference update with no local improvement covector is outside the current formulation. Likewise, closed API-only generators are outside our target setting because joint generator post-training itself already assumes access to internal transitions and parameters.

D.4 Benchmark protocol

VBench (Huang et al., 2024) and VBench-2.0 (Zheng et al., 2025) use their official aggregate/component definitions. T2V-CompBench (Sun et al., 2025) uses its official compositional dimensions and reported aggregate. StoryEval (Wang et al., 2025) follows the same official event-completion protocol for every generator updater and backbone. Every checkpoint uses identical resolution, duration, scheduler steps, negative-prompt convention, and three fixed generation seeds per benchmark request. The three scores for a request are averaged first; official benchmark aggregation is then applied to those request-level means. This preserves matched stochasticity while preventing seed count from being treated as sample size.

We distinguish three sources of uncertainty: training-run variation, within-checkpoint generation randomness, and request sampling. The headline hierarchical-credit-versus-Full comparison uses three independent training runs; the functional-coordination factorial uses the same three-run protocol, and other trainable baselines use two independent runs. Confidence intervals for the headline increment are computed on paired run-level deltas using a Student-t interval; request-level paired bootstrap intervals are retained as a within-run sensitivity analysis. Generation seeds are averaged within each checkpoint and are never treated as independent training replicates. The same checkpoint and seed definitions are used across all controlled comparisons.

D.5 Baseline implementation

All trainable baselines draw from the same 8,000-request training pool and use the same number of request draws and decoded videos unless a comparison explicitly targets compute. The frozen official LTX-2.5 prompt enhancer uses the reference LTX-2.5 enhancement path (Lightricks, 2026): `google/gemma-4-E2B-it` with `LTX2_5_T2V_DEFAULT_SYSTEM_PROMPT`, greedy decoding (`do_sample=False`), `no_repeat_ngram_size=5`, and at most 600 new tokens. Prompt-only updates Qwen3.5-9B while freezing LTX-2.5; the Flow-GRPO generator-only row keeps the original request as its text interface and is included as a single-side reference rather than as the matched joint-training control. Prompt-A-Video (Ji et al., 2025) and VPO (Cheng et al., 2025) use their published prompt-side objectives within the common Qwen3.5-9B prompt optimizer, training pool, LTX-2.5 generator, and evaluation harness. Because their original backbones and training stacks differ, the reported values are controlled common-harness results rather than copied published absolute scores.

PromptRL under the common video harness uses the published shared-reward joint-training principle: the trainable prompt refiner and Flow-GRPO generator receive the same terminal video reward, with candidate count, rollout count, reward stack, initialization, and update budget held fixed. UniGRPO adds group-relative normalization for the text policy and generator branch within the same joint rollout groups, retaining the published unified sparse-terminal-reward idea while keeping our LTX-2.5

backbone and reward stack fixed. We use these rows to compare joint-optimization principles within one harness, not as claims of exact reproduction of the original image-generation systems (Wang et al., 2026a; Liu et al., 2026). Matched Global-LOO then trains the same two modules on the same $K \times M$ grid with the global leave-one-out definition in Appendix A.2; hierarchical credit changes only the credit hierarchy, and Full adds functional prompt shaping and generator routing.

CAPE-T2V (Jia et al., 2026) is the closest T2V-specific two-sided system in our comparison set, but it addresses the prompt–caption distribution gap through sequential Anchored-PE training followed by DiT adaptation on a separately constructed 54k video-derived caption set. Its released LTX pipeline targets LTX-2.3, whereas our mechanism-controlled study uses online joint reward post-training on LTX-2.5 with no auxiliary video-caption supervision. We therefore use CAPE-T2V for external positioning rather than treating its public absolute scores as a matched-budget ranking.

Table 7: Positioning of two-sided adaptation systems. The first two rows are controlled LTX-2.5 implementations used numerically in Table 1; CAPE-T2V is a heterogeneous external reference and is not mixed into the matched-budget ranking.

System	Prompt side	Generator side	Additional supervision	Training form
PromptRL	trainable LM refiner	Flow-GRPO	none	online shared-terminal-reward joint update
UniGRPO	trainable text policy	Flow-GRPO	none	online group-relative joint policy optimization
CAPE-T2V	Anchored PE	DiT fine-tuning	54k video-derived captions	sequential two-sided conditioning alignment
FLOWCRED	Qwen3.5-9B policy	Flow-GRPO	none	online hierarchical credit + functional coordination

Wrong-query controls are matched to the target request by compositional category, raw-prompt length, enhanced-prompt length, candidate count, and hierarchical-credit difficulty bin. Random spans are rank matched after Fisher whitening and sample response norms from the empirical true-span distribution. Uniform shrink chooses its scalar on the development split.

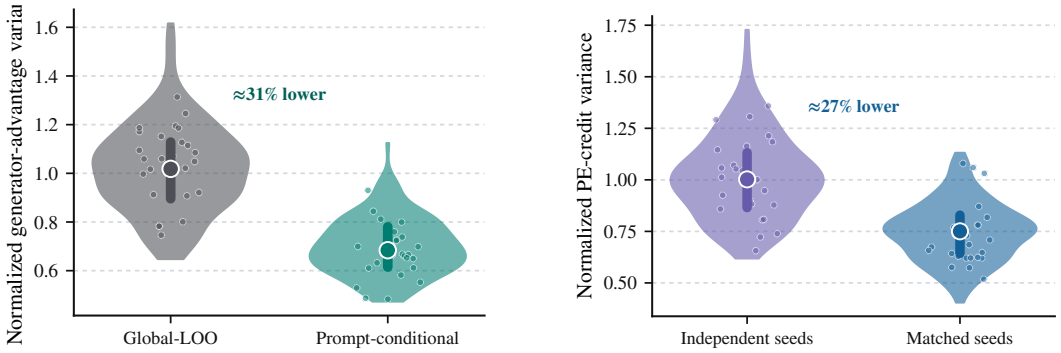
D.6 AGGREGATE BENCHMARK SUMMARY

Table 8: Aggregate benchmark status for the principal LTX-2.5 systems under the unified evaluation harness.

Benchmark	Metric	Base	Prompt-only	Global-LOO	Hier. credit	Full
VBench	Total	78.14	86.97	86.49	86.91	**87.56**
VBench-2.0	Overall	54.52	62.84	63.41	63.92	**64.73**
T2V-CompBench	Overall	46.52	60.28	61.47	62.63	**64.46**
StoryEval	Overall	37.88	54.37	55.58	56.71	**58.08**

E HIERARCHICAL CREDIT ESTIMATOR EFFICIENCY

Hierarchical credit makes two testable statistical predictions: prompt-conditioned generator baselines remove irrelevant prompt-to-prompt variation, while matched seeds reduce prompt-policy contrast variance when same-seed rewards are positively correlated. Both effects are measured directly rather than inferred from endpoint quality.



(a) Generator advantage variance: global leave-one-out versus prompt-conditional residual.

(b) Prompt candidate-credit variance: independent versus matched seeds at equal video count.

Figure 6: Estimator-efficiency diagnostics across held-out minibatches: prompt-conditioned generator baselines reduce generator-advantage variance, and matched seeds reduce prompt candidate-credit variance at equal video count.

Generator-advantage variance is reduced by about 31%, and matched seeds reduce prompt-credit variance by about 27%. The matched-seed ablation keeps total decoded videos fixed and yields 62.63 T2V-CompBench versus 62.21 with independently sampled seeds.

F Mechanism Diagnostics

Table 9: Functional-coordination factorial ablation. Mean (SD) across the same three-run protocol used for the HC–Full comparison.

Prompt shaping	Generator routing	VBench↑	StoryEval↑	T2V-CompBench↑
–	–	86.91 (0.14)	56.71 (0.27)	62.63 (0.32)
✓	–	87.20 (0.15)	57.36 (0.28)	63.40 (0.34)
–	✓	87.29 (0.13)	57.18 (0.26)	63.67 (0.31)
✓	✓	**87.56 (0.12)**	**58.08 (0.24)**	**64.46 (0.29)**

F.1 Realized-update audit

For cached states, let $\Delta\mu_t^{\text{real}} = \mu_{\theta^+}(s_t, p_k) - \mu_\theta(s_t, p_k)$. Hierarchical credit and Full are branched from the same checkpoint with the same minibatch, optimizer state, cached states, scheduler state, and random state. We report

$$\rho_{\text{span}}^{\text{real}} = \frac{\|P_{\lambda,t}\Delta\mu_t^{\text{real}}\|_{F_t}^2}{\|\Delta\mu_t^{\text{real}}\|_{F_t}^2 + \epsilon}, \tag{31}$$

together with $\|\Delta\mu_{\text{Full}}^{\text{real}}\|_F / \|\Delta\mu_{\text{HC}}^{\text{real}}\|_F$. The realized comparison is 0.487→0.264 for supported overlap and 0.931 for the update-norm ratio. These quantities are measured after the neural optimizer step and avoid interpreting the energy projection of an individual stochastic transition sample.

F.2 Prompt-policy shaping and fresh-seed utility

The mechanism set contains 360 requests with four candidates each. Scores use the matched $M = 4$ training seeds and utility uses eight disjoint fresh seeds; bootstrap intervals resample requests while preserving candidate groups. An 80-request subset with 16 fresh seeds checks ranking stability. Controls include response norm, functional-only, reward-only, cosine-functional, and the default trust-clipped directional score.

F.3 True, random, wrong-query, and generator-specific geometry

Wrong-query spans are matched by compositional category, prompt length, candidate count, and hierarchical-credit difficulty. Evaluation uses independent $B_{\text{eval}} = 12$ alternatives that never enter

the training gate. Generator specificity scores identical candidate sets with Base, reward-trained, and Full snapshots and evaluates fresh-seed utility under each rollout generator.

G Prompt Integrity and Data-Audit Controls

Because Prompt-only already produces large absolute gains, all evaluator inputs remain the original request u, and the paper separately audits verbosity, semantic drift, and train/evaluation leakage.

Table 10: Prompt-integrity diagnostics. Hierarchical credit and Full use nearly identical prompt lengths and semantic-filter behavior.

Diagnostic	Raw request	Prompt-only	HC	Full
Median tokens	18	111	117	120
P90 tokens	31	149	153	157
Intent-filter rejection	–	4.3%	4.7%	4.5%
Blind semantic drift ($N = 200$)	–	7/200	5/200	5/200

On a token-length-matched subset, Full still gains +0.61 VBench, +1.22 StoryEval, and +1.59 T2V-CompBench over hierarchical credit. Training/evaluation separation uses normalized-string matching plus lexical and embedding nearest-neighbor retrieval with thresholds fixed before benchmark reporting.

H Complete Mechanism and Control Audit

For compactness in the ablation plots and tables, "HC" denotes the hierarchical-credit system without functional coordination.

Table 11: Compact mechanism and routing audit. Panel A evaluates prompt-candidate scores on disjoint fresh seeds; Panel B tests matched-budget routing alternatives.

A. Fresh-seed prompt-candidate ranking

Candidate score	Spearman ρ	Top-1
Response norm only	0.313 [0.251, 0.374]	50.3% [46.1,54.4]
Functional only	0.421 [0.369, 0.472]	55.8% [51.6,60.0]
Reward only	0.447 [0.398, 0.497]	59.2% [55.0,63.2]
Reward + cosine functional	0.569 [0.519, 0.618]	67.2% [63.2,71.0]
Reward + directional functional	**0.612 [0.568, 0.655]**	**71.7% [67.8,75.4]**

B. Matched-budget routing controls

Variant	T2V-CompBench	Δ vs. HC
Hierarchical credit	62.63	0.00
Rank-matched random span	62.55	-0.08
Matched wrong-query span	62.69	+0.06
Uniform gradient shrink	62.81	+0.18
No cross-fit	63.44	+0.81
No reliability gate	63.67	+1.04
Full	**64.46**	**+1.83**

I Generalization and Compute

I.1 Backbone and generator-updater transfer

The Hunyuan experiment preserves candidate construction, matched estimators, executed-action reference, and allocation logic while substituting backbone-native transition quantities. RVM and Diffusion-DRF comparisons are likewise within-updater: hierarchy-only versus the same system plus functional coordination under an identical StoryEval/T2V-CompBench harness. We therefore claim positive within-setting increments, not cross-optimizer rankings. Under the same VBench-2.0 protocol, the measured raw-prompt Overall scores are 54.79 for HunyuanVideo-1.5 and 54.52 for LTX-2.5.

Table 12: Transfer comparisons across a second backbone and two alternative generator updaters, including differentiable reward flow. Absolute scores are shown within each matched setting rather than used to rank heterogeneous optimizers.

Method / within-setting contrast	VBench	T2V-CompBench	StoryEval
HunyuanVideo-1.5 / Flow-GRPO			
Matched Global-LOO	84.67	55.34	55.12
Hier. credit	84.96	55.89	56.04
Full	**85.53**	**57.11**	**57.13**
LTX-2.5 / RVM			
Hier. credit	86.39	61.28	56.18
Full	**86.91**	**62.57**	**57.33**
LTX-2.5 / Diffusion-DRF			
Hier. credit	87.08	63.16	57.24
Full	**87.49**	**64.13**	**58.03**

I.2 Local-compute accounting

Full keeps the decoded-video and reward-model-call budget identical to hierarchical credit. Under the default $K = 4$ and $B_{\text{train}} = 6$, it additionally samples 96 accepted support prompts per request (six at the routed state of each of 16 trajectories) and performs 144 single-state generator transition forwards: 48 for the $K - 1 = 3$ rollout-candidate responses and 96 for the support probes. These operations do not decode additional videos or invoke the video reward stack. The same frozen intent checker screens support prompts; extra proposal/checker calls caused by rejection resampling are data-dependent and are therefore not included in the deterministic count. End-to-end hardware cost depends on the generator implementation and prompt-policy sampling stack, so we report deterministic operation counts rather than presenting a universal accelerator-hour ratio.

Table 13: Deterministic per-request operation counts for the primary LTX setting.

Quantity	Hierarchical credit	Full
Decoded videos / request	16	16
Reward-model calls / request	16	16
Accepted support prompts / request	0	96
Same-state transition probe forwards	0	144

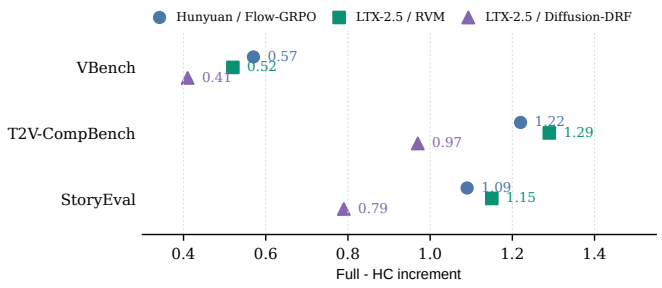


Figure 7: Within-setting transfer of the Full−hierarchical-credit increment across a second backbone and two alternative generator updaters.

J Robustness and Replication

We keep robustness reporting focused on two questions that materially affect the main claim: whether the method remains useful when the decoded rollout budget changes, and whether the hierarchical-

credit-to-Full increment repeats across independent training runs. Hyperparameters are fixed before the reported test evaluation; benchmark test scores are not used for selection.

Table 14: Decoded-rollout budget study with $K = 4$. Hierarchical credit requires $M \geq 2$ because the prompt-conditioned leave-one-seed-out residual is undefined for a single trajectory.

M	Videos/request	PromptRL	Global-LOO	Hier. credit	Full	Full−HC
2	8	59.74	60.21	61.11	**63.18**	+2.07
4	16	61.06	61.47	62.63	**64.46**	+1.83
8	32	62.11	62.38	63.34	**64.87**	+1.53

All methods improve with more terminal evidence, while the Full-minus-HC increment remains positive across the three budgets. We treat M as a compute axis rather than a free hyperparameter improvement.

The main hierarchical-credit-versus-Full comparison is repeated for three independent training runs. Run-level deltas, not generation-seed cells, are the outer replication units.

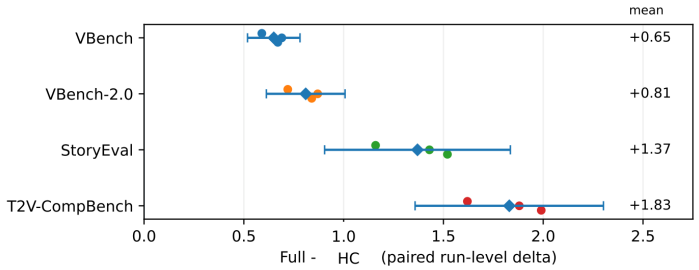


Figure 8: Three-run replication. Small points are paired training-run deltas; diamonds show means and horizontal intervals show Student-t 95% intervals over the three run-level differences.

Table 15: Raw Full−hierarchical-credit deltas for three independent training replicates.

Metric	Rep. A	Rep. B	Rep. C
VBench	0.59	0.69	0.67
VBench-2.0	0.72	0.87	0.84
T2V-CompBench	1.62	1.88	1.99
StoryEval	1.16	1.43	1.52

The corresponding means are 0.65 VBench, 0.81 VBench-2.0, 1.83 T2V-CompBench, and 1.37 StoryEval. With only three outer replicates, we use these intervals descriptively and do not turn them into stronger significance claims.

K Realized Cross-Module Coordination Audit

The response span establishes local conditioning capacity; this audit asks whether the trained modules subsequently move in the complementary pattern predicted by functional coordination. For each held-out routed state, we cache the pre-update behavior generator and estimate the prompt-policy-induced displacement

$$\Delta\mu_{\mathrm{P}} = \mathbb{E}_{p\sim\pi_{\phi^+}}[\mu_{\theta^-}(s,p)] - \mathbb{E}_{p\sim\pi_{\phi}}[\mu_{\theta^-}(s,p)].$$

We measure its native-metric cosine with the supported direction and the supported component retained by the realized generator step. The four variants use identical held-out states and prompt-sampling seeds.

Table 16: Realized cross-module coordination on held-out states.

Training variant	$\cos_F(\Delta\mu_P, P_\lambda g^R)$	Generator supported overlap ↓
Hierarchical credit	0.29	0.487
Prompt shaping only	0.55	0.471
Generator routing only	0.30	0.281
Full	**0.61**	**0.264**

The factorial pattern is consistent with the intended division of labor: prompt shaping mainly changes the upstream displacement, generator routing mainly reduces the downstream component lying in the response span, and Full exhibits both effects. We treat this as a descriptive consistency audit rather than a causal decomposition.

L Human Evaluation and Qualitative Cases

Expanded qualitative cases. Figure 9 adds four benchmark examples chosen to broaden the qualitative coverage beyond the two main-text cases. The panel spans numeracy/composition, multi-person interaction, spatial relation, and a coupled appearance transition; rows again compare Base, VPO, and FlowCred using representative temporal frames.



Figure 9: **Additional qualitative comparisons.** Four benchmark cases covering numeracy, interaction, spatial relation, and synchronized appearance change. Rows show Base, VPO, and FlowCred.

(a) Numeracy and composition. The key requirement is exact cardinality together with a coherent shared-meal scene. Base over-populates the table, so the generated composition no longer matches the requested five-person configuration. VPO moves in the opposite direction and under-populates the scene. FlowCred maintains five visible participants across the representative frames while keeping them organized around one table, preserving both the requested count and the intended conversational meal context rather than satisfying one at the expense of the other.

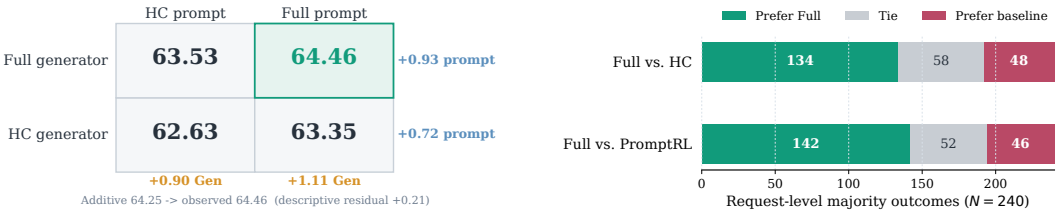
(b) Protagonist-focused interaction. This case tests whether the model can isolate the intended pair and carry the reunion interaction through to a clear embrace. Base disperses the event into a crowded group hug, making the two target friends difficult to identify consistently. VPO produces an embrace but shows weaker protagonist focus across the sequence. FlowCred gives the clearest two-person progression from greeting to contact to hug, while maintaining the same pair more consistently through the action.

(c) Spatial relation. The decisive constraint is the word *above*, so object presence alone is insufficient. Base and VPO both render a donut and a handbag but fail to maintain the requested relative

placement in a stable, unambiguous form. FLOWCRED consistently places the donut above the handbag across the representative frames, turning an otherwise plausible pair of objects into the correct relational composition. This mirrors the larger gains on compositional evaluation, where satisfying relations rather than merely rendering entities is the harder requirement.

(d) Coupled color transition. The final case requires a coordinated temporal state change: both the chameleon and the neon sign should progress from green toward red. Base does not complete the requested transformation, leaving the appearance change partial. VPO reaches a red state but does so abruptly, with a less natural visual progression and a conspicuous change in camera behavior. FLOWCRED produces the most gradual and complete transition: the two scene elements evolve coherently toward red while the framing remains substantially more stable, making the requested transformation easier to follow as a continuous event rather than a sudden endpoint switch.

Blinded human evaluation. We sample 240 requests stratified across compositional and multi-event categories. Three blinded annotators receive the original request and anonymized matched-seed videos with randomized left/right placement and score faithfulness, event/relation completion, motion/visual quality, and overall preference. Raw request-level proportions retain ties; the two primary significance tests condition on non-tie Full-versus-baseline outcomes and use exact binomial tests with Holm correction. Per-dimension analyses are descriptive, and Fleiss' κ reports agreement.



(a) Frozen-module 2 × 2 cross-swap. Either Full-trained module improves the HC pair when substituted alone; the +0.21 residual is descriptive, not a causal synergy claim.

(b) Blinded request-level human preference with ties retained.

Figure 10: **External validation beyond automatic benchmarks.** Frozen-module substitution attributes gains to both trained modules, while human judgments independently prefer Full over HC and PromptRL. Per-dimension vote rates are reported in Table 17.

Table 17: **Blinded matched-seed human evaluation.** Panel A reports request-level majority outcomes and non-tie Full win rates with Wilson 95% intervals; Panel B reports per-dimension Full preference rates with ties retained.

A. Primary request-level preference

Comparison	Prefer Full	Tie	Prefer baseline	Non-tie Full win	Holm p	Fleiss κ
Full vs. HC	134	58	48	73.6% [66.8, 79.5]	1.34×10^{-10}	0.51
Full vs. PromptRL	142	52	46	75.5% [68.9, 81.1]	2.72×10^{-12}	0.53

B. Per-dimension Full preference

Comparison	Dimension	Prefer Full	95% CI
Full vs. HC	Faithfulness	128/240 (53.3%)	[47.0, 59.5]
	Composition/event	151/240 (62.9%)	[56.6, 68.8]
	Motion/visual	122/240 (50.8%)	[44.5, 57.1]
	Overall	134/240 (55.8%)	[49.5, 62.0]
Full vs. PromptRL	Faithfulness	132/240 (55.0%)	[48.7, 61.2]
	Composition/event	158/240 (65.8%)	[59.6, 71.5]
	Motion/visual	127/240 (52.9%)	[46.6, 59.1]
	Overall	142/240 (59.2%)	[52.9, 65.2]

Preference pattern. The request-level majority analysis favors Full in both primary comparisons. Against HC, Full wins 134 requests versus 48 baseline wins, with 58 ties; conditioning on non-ties gives a 73.6% Full win rate. Against PromptRL, the corresponding counts are 142 / 52 / 46 and a

75.5% non-tie win rate. The Holm-corrected exact tests are significant in both comparisons, while Fleiss' κ values of 0.51 and 0.53 indicate moderate agreement rather than near-deterministic annotator consensus.

Where human gains concentrate. The largest descriptive gap occurs on composition/event completion: Full is preferred on 62.9% of requests against HC and 65.8% against PromptRL. Faithfulness improves more modestly (53.3% and 55.0%), and motion/visual preference remains closer to parity (50.8% and 52.9%). This pattern is consistent with the automatic results, where the largest gains appear on compositional and multi-event benchmarks, while also showing that the preference advantage is not driven solely by a generic visual-quality shift.

Frozen-module attribution. The 2×2 post-training cross-swap uses HC/Full prompt optimizers $\times$ HC/Full generators and freezes all modules during the audit. Substituting the Full generator into the HC pair contributes +0.90 T2V-CompBench, while substituting the Full prompt optimizer contributes +0.72; with the complementary module already Full-trained, the corresponding increments are +1.11 and +0.93. Starting from 62.63, the HC-conditioned marginal effects predict 64.25 additively, and the Full pair reaches 64.46, leaving only a descriptive +0.21 residual. The audit therefore attributes measurable gains to both trained sides without requiring a strong interaction effect between them.