TOKEN-LEVEL VIDEO REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

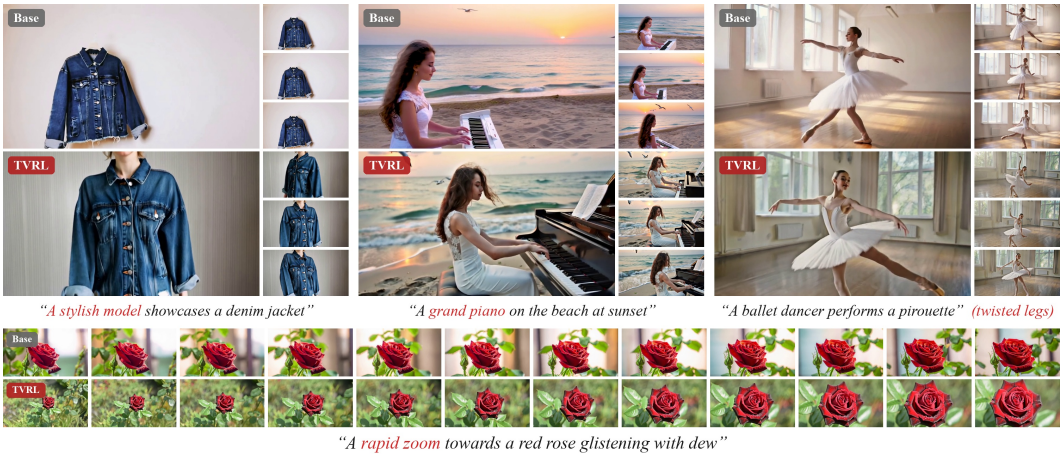

Figure 1: **Text-to-video RL post-training with TVRL.** Compared with the base model (top rows), TVRL (bottom rows) improves subject presence, object fidelity, body plausibility, and camera-motion control; red marks the prompt phrase to check.

ABSTRACT

Reinforcement learning (RL) for video generation usually assigns one scalar reward to an entire sampled video. Yet a video is not uniformly flawed: some visual tokens may already satisfy the prompt, whereas others require correction. A scalar reward cannot localize errors, causing optimization to perturb satisfactory tokens while under-targeting the tokens that actually need to change. We introduce Token-Level Video Reinforcement Learning, TVRL, a framework that derives token-level credit from the reward being optimized. Our key insight is that the answer likelihood of a frozen vision-language model provides both signals: its outputs contribute to the video-level reward, while magnitudes of its video-input gradients reveal which generated video tokens most affect that score. We instantiate TVRL in Group Relative Policy Optimization by averaging prompt-derived question rewards into one group-relative advantage and using detached, question-conditioned token-credit maps to reweight dense denoising-transition log-probabilities inside the clipped policy ratio. On VBench-2.0, TVRL achieves an Overall score of 57.69, outperforming the base model by 3.60 points. TVRL also improves matched GRPO baselines across three SDE samplers (SAGE, Flow, and Dance) by 2.68–3.15 points and across four reward models (VideoAlign, VideoScore2, UnifiedReward2, and Qwen3.5-9B) by 1.33–3.15 points.

1 INTRODUCTION

Text-to-video generation (Wan et al., 2025; Kong et al., 2024; Wu et al., 2025a; Polyak et al., 2024; Lightricks, 2025) has advanced rapidly with diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022; Ho et al., 2022) and flow-matching (Lipman et al., 2023) models. As base models improve, post-training that aligns generated videos with prompts, human preferences, and evaluation criteria has become increasingly important (Xu et al., 2023; Liu et al., 2025b; Xu et al., 2024). Inspired by reinforcement learning for language models (Ouyang et al., 2022; Shao et al., 2024; DeepSeek-AI, 2025), recent work adapts Group Relative Policy Optimization (GRPO) (Shao

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.

et al., 2024) to video generation (Liu et al., 2025a; Xue et al., 2025; Zheng et al., 2026) by converting the deterministic denoising ODE into an SDE sampler, which defines a stochastic policy over latent denoising trajectories. For each prompt, GRPO samples a group of videos, scores each completed trajectory, and updates the generator with group-relative advantages, with no separate value model.

Despite this progress, video RL faces a fundamental credit-assignment bottleneck. Training must determine not only whether one trajectory is better than another, but also when and where the reward-relevant evidence appears. A video-level scalar reward can rank completed denoising trajectories, but it cannot indicate which video tokens are responsible for success or failure (Wu et al., 2025b; Wang et al., 2026; Liang et al., 2026) (Figure 2). This limitation matters because the evidence for many prompt requirements is confined to a small region or a few frames, for example, whether the requested grand piano is rendered as a piano rather than a generic keyboard (Figure 1), whether a watch case is rose gold, whether the described fireflies appear in the scene at all, or whether the camera pans to the right instead of staying static. Broadcasting the same scalar advantage to every latent token therefore spends update capacity on regions that have nothing to do with the evaluated error. Recent work separates the *video-level reward* from *token credit* (Ni et al., 2026), but derives credits from perceptual features independent of rewards.

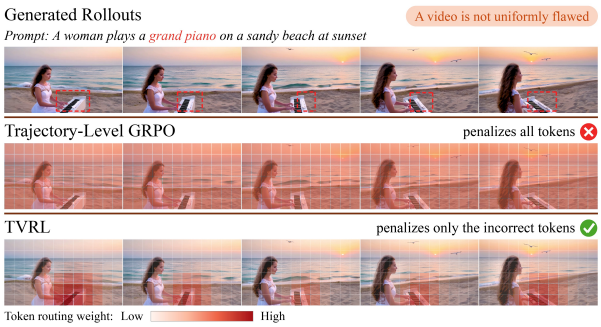

Figure 2: **Video-level reward, token-level credit.** Video GRPO (Zheng et al., 2026) broadcasts one scalar advantage to every video token. TVRL provides token-level credit to acknowledge or penalize relevant tokens.

The missing quantity is a routing signal tied to the reward itself. Score gradients provide such a signal because they measure how sensitive an evaluator's score is to local changes in the video-frame input (Xu et al., 2023; Clark et al., 2023; Prabhudesai et al., 2023; 2024). Prior differentiable reward methods typically use these gradients to update the generator directly. In GRPO, we use them differently: after detachment, their magnitudes weight denoising log-probabilities, while the group-relative advantage remains the scalar RL signal. This preserves the clipped policy-gradient structure of GRPO and avoids backpropagating the reward loss through the generator. The remaining requirement is semantic alignment: the score being differentiated must represent the requirement whose credit we want to localize.

This score–credit alignment makes the choice of evaluator interface central. Custom video reward models provide differentiable scalar feedback, but a single scalar head does not expose which semantic requirement its gradient should explain (Wang et al., 2026). We therefore use an evaluator whose score is tied to an explicit video-grounded question, so the reward and the sensitivity map are computed from the same semantic check. Pretrained VLMs (Qwen Team, 2026; Bai et al., 2025) offer exactly such an interface: prompt following, physical plausibility, object interaction, and temporal consistency can all be phrased as video-grounded questions, and the teacher-forced likelihood of a reference answer measures whether each check is satisfied. Because every question–answer likelihood is differentiable with respect to the VLM's video-frame input, a single check provides both a scalar contribution to the rollout objective and a question-conditioned sensitivity map for routing its update. Using a frozen, off-the-shelf evaluator also removes the need for reward-model training and per-check preference labels (Furuta et al., 2024). This is the central idea of our work: the same semantic check that determines *whether* a prompt requirement is satisfied also estimates *where* the resulting policy update should be applied.

Building on this observation, we propose TVRL (Token-Level Video Reinforcement Learning), a training framework in which a video policy is evaluated by video-level rewards but updated through explicit token-level credit; the name does not assume that video tokens are discrete actions. We instantiate the framework for text-to-video diffusion with a clipped group-relative objective. For each prompt, we construct binary video-grounded checks and score generated videos by the teacher-forced VLM likelihood of the reference answers. Each check yields two coupled quantities from the same scalar: a question reward given by the likelihood value, and a detached token-credit proxy given by the magnitude of its video-frame-input gradient. The latter measures local sensitivity rather

than signed or causal credit. We average the question rewards before group-relative normalization to obtain a single rollout-level advantage shared across checks, and we use each question-conditioned credit map to reweight the dense denoising-transition log-probabilities inside the clipped GRPO surrogate. The shared advantage sets the direction and strength of the update, while the credit maps determine its spatio-temporal allocation. This design preserves group-relative rollout ranking and ratio clipping, and it never backpropagates reward gradients through the generator. Averaging several checks also reduces reliance on any single question. Our **contributions** are:

- We derive token-level credit for video GRPO from the reward itself: a frozen VLM supplies both a question reward and a detached, question-conditioned sensitivity map, requiring neither task-specific reward-model training nor dense credit labels.
- We develop a shared-advantage, localized-routing objective that aggregates question rewards into one rollout-level advantage and uses the credit maps to reweight dense denoising-transition log-probabilities within the clipped GRPO ratio. Under the baseline's log-probability reduction, uniform weights recover scalar-reward GRPO.
- With the SAGE sampler, TVRL improves VBench-2.0 Overall over the corresponding GRPO baselines by +1.36, +1.33, +1.85, and +3.15 points with VideoAlign, VideoScore2, UnifiedReward2, and Qwen3.5-9B rewards, respectively. With the Qwen3.5-9B reward fixed, it also improves over GRPO under the SAGE, Flow, and Dance samplers by +3.15, +2.92, and +2.68 points. Ablations cover credit granularity and the choice of VLM critic.

2 Related Work

Token-level credit for video optimization. Alignment of video generators has increasingly moved beyond a single global score toward feedback structured by space, time, or evaluation criterion. Differentiable reward optimization obtains such feedback by backpropagating reward gradients through the decoder and the denoising chain (Clark et al., 2023; Prabhudesai et al., 2023; 2024; Yuan et al., 2024; Wang et al., 2026). These gradients encode how sensitive the score is to local video evidence, but using them to update the generator directly requires keeping the reward model, the decoder, and the multi-step denoising graph in memory, and repeated Jacobian products along the stochastic chain can amplify unstable local gradients. Other dense-feedback methods expose localized failures through temporal preference labels, dense captions, rich human or AI feedback, or patch-level reward models (Wu et al., 2025b; Krishna et al., 2017; Xiong et al., 2024; Liang et al., 2024; Liu et al., 2025b; Wang et al., 2025a), at the cost of additional annotation or a separately trained localized reward model. Structured alignment systems (Liang et al., 2026) and ViPO (Ni et al., 2026) redistribute a scalar evaluation or GRPO advantage to finer units, yet their allocation maps come from perceptual priors rather than from the check that produced the reward. Recent VLM-based video evaluators instead express prompt requirements as semantic questions rather than task-specific reward heads (Qwen Team, 2026; He et al., 2025; Wang et al., 2025b). TVRL uses this interface to couple reward evaluation with token-level credit: teacher-forced answer likelihoods score prompt-level checks, and detached video-frame-input gradients weight the denoising log-probabilities within GRPO. Unlike Diffusion-DRF (Wang et al., 2026), which also uses structured VLM feedback but backpropagates it through the generator, the VLM gradient here stops at the credit map; unlike scalar GRPO, it changes where the update is applied; and unlike dense-feedback methods, it needs no dense labels or localized reward model.

GRPO for video generation. GRPO removes the need for a learned value model by sampling multiple outputs for the same prompt and normalizing their rewards within the group, which makes it attractive for large video generators (Shao et al., 2024). Recent video GRPO methods adapt this formulation to diffusion and flow models by treating denoising as a stochastic policy and defining transition log-probabilities along sampled latent trajectories. Flow-GRPO (Liu et al., 2025a), DanceGRPO (Xue et al., 2025), and SAGE-GRPO (Zheng et al., 2026) instantiate this framework with different stochastic denoising samplers. All of them optimize a clipped policy-gradient objective, but with a scalar reward they assign one terminal reward to each rollout and apply the resulting group-relative advantage uniformly across denoising steps and latent video tokens. Such a reward indicates whether a video is better than the others in its prompt group, not where the update should concentrate. Our video instantiation of TVRL is complementary to these samplers: it retains the

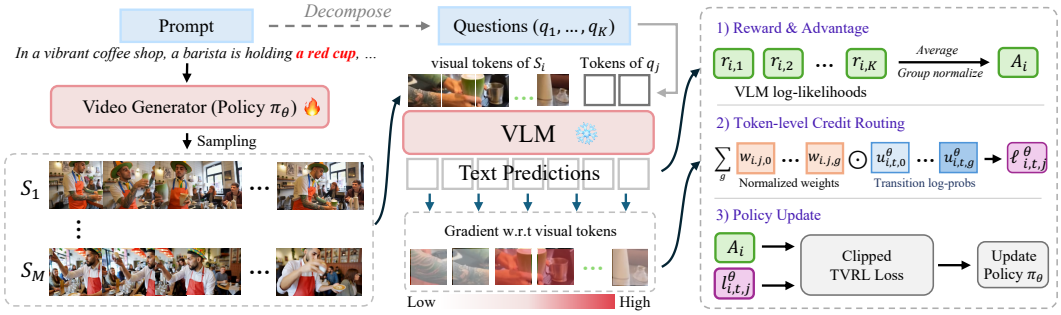

Figure 3: Overview of TVRL. Teacher-forced likelihoods for K prompt-derived checks are averaged within each of M rollouts and normalized across the rollout group into one advantage per video. Gradients of the same likelihoods yield detached, question-conditioned weights that route dense denoising log-probabilities inside the clipped GRPO update.

clipped group-relative update and replaces uniform assignment with VLM-gradient token credit that reweights the dense denoising log-probabilities before the policy ratio is formed.

3 METHOD

In *token-level video reinforcement learning*, a video-level reward decides whether a rollout is reinforced or suppressed, and an explicit credit distribution decides how that update is allocated over video tokens. With rollout advantage A_i, local policy score $u^\theta_{i,t,g}$ of video unit g, and credit $w_{i,g} \geq 0$ with $\sum_g w_{i,g} = 1$, the unclipped surrogate gradient at the old policy is $-A_i \sum_g w_{i,g} \nabla_\theta u^\theta_{i,t,g}$. Uniform credit recovers scalar GRPO; ViPO (Ni et al., 2026) derives w from perceptual features, whereas we derive it from the reward's own input gradient.

We instantiate this for a text-to-video diffusion policy π_θ with a frozen vision-language model (VLM). Given a prompt c, the policy samples a denoising trajectory $\tau = (\mathbf{z}_T, \ldots, \mathbf{z}_0)$, a frozen decoder produces the video $\mathbf{v} = D(\mathbf{z}_0)$, and the VLM scores sampled frames by teacher-forced question answering. The question rewards form one group-relative advantage per rollout, and their detached VLM gradients only route it over latent units inside the clipped GRPO objective; no VLM gradient reaches the generator (Figure 3).

3.1 POLICY AND VIDEO-GROUNDED CHECKS

A pretrained diffusion transformer (Peebles & Xie, 2023; Wan et al., 2025) defines stochastic transitions $\pi_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, t, c)$. During GRPO rollouts, the SDE sampler perturbs the transition mean $\boldsymbol{\mu}^{\text{sde}}_\theta$ with noise schedule $\sigma_{\eta,t}$:

$$\begin{aligned} \mathbf{z}_{t-1} &\sim \mathcal{N}\big(\boldsymbol{\mu}^{\text{sde}}_\theta(\mathbf{z}_t, t, c), \sigma^2_{\eta,t} I\big)\,, \\ \log \pi_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, t, c) &= -\frac{\|\mathbf{z}_{t-1} - \boldsymbol{\mu}^{\text{sde}}_\theta(\mathbf{z}_t, t, c)\|_2^2}{2\sigma^2_{\eta,t}} + \text{const.} \end{aligned} \tag{1}$$

This transition yields an elementwise latent log-probability map whose gradients reach the generator through the transition mean. Different SDE samplers change the mean and the noise schedule, but TVRL requires only that the dense log-probabilities can be reduced over a chosen set of credit units. Appendix B.1 gives the full scheduler transition.

We decompose each prompt offline into K atomic yes/no checks $\mathcal{Q}(c) = \{q_j(c)\}_{j=1}^K$. The decomposition model sees only the prompt, and each check targets a single requirement directly verifiable from the generated video, such as a subject, attribute, action, relation, scene element, or camera cue. At reward time, all checks and their fixed reference answers are placed in a structured schema and scored in a single teacher-forced VLM pass (templates in Appendices B.5 and B.6).

3.2 TEACHER-FORCED REWARD AND TOKEN CREDIT

Let p_ϕ denote the frozen VLM. For rollout i and check j, the question reward is the average log-probability of the fixed reference answer over its answer-token positions $\mathcal{A}_{i,j}$:

$$R_{i,j} = \frac{1}{|\mathcal{A}_{i,j}|} \sum_{\ell \in \mathcal{A}_{i,j}} \log p_\phi\big(a_{i,j}^{\ell} \mid \mathbf{x}_i, a_i^{<\ell}\big)\,. \tag{2}$$

We differentiate the same score with respect to the processed video-frame input $\boldsymbol{\xi}_i$. For a credit granularity Γ with latent credit units $\mathcal{G}_\Gamma$, the gradient magnitude is aggregated, aligned with these units, and normalized:

$$\begin{aligned} \mathbf{g}_{i,j} &= \nabla_{\boldsymbol{\xi}_i} R_{i,j}, & \tilde{w}_{i,j,g} &= \mathrm{Agg}_\Gamma(|\mathbf{g}_{i,j}|)_g, \\ w_{i,j,g} &= \frac{\tilde{w}_{i,j,g}}{\max(\sum_{g' \in \mathcal{G}_\Gamma} \tilde{w}_{i,j,g'}, \epsilon)}, & g &\in \mathcal{G}_\Gamma. \end{aligned} \tag{3}$$

Depending on Γ, Agg_Γ produces frame-level or spatially smoothed credit, falling back to uniform weights for invalid maps; granularity trades localization against noise (Appendix B.7).

We call $|\mathbf{g}_{i,j}|$ the *reward-sensitivity map* of check j: to first order, it measures how much the check's reward changes under a small perturbation of each video token; it is local and model-dependent, not a causal attribution. Unlike bottom-up saliency or foreground maps, it depends on the question and its reference answer rather than on the video alone: any score-independent map has zero question-swap interaction by construction, whereas ours is positive and consistent in sign (Appendix A.6). We reserve *token credit* for the normalized routing weights $w_{i,j,g}$.

3.3 SHARED ROLLOUT ADVANTAGE

For each prompt, the old policy samples a group of M rollouts. We first average the question rewards within each rollout and then normalize across the prompt group:

$$\bar{R}_i = \frac{1}{K} \sum_{j=1}^{K} R_{i,j}, \qquad A_i^{\mathrm{tvrl}} = \frac{\bar{R}_i - \mu_G}{\max(\sigma_G, \epsilon)}, \tag{4}$$

where μ_G and σ_G are the mean and population standard deviation over the M rollouts. This order preserves the video-level ranking and avoids separate normalization statistics for each check. All checks share A_i^{tvrl} and route it through their own maps; a failed check can therefore receive a positive shared advantage. The maps allocate the video-level reward rather than supervise each question.

3.4 TOKEN-LEVEL POLICY UPDATE

Let Ω_g denote the entries of the dense transition log-probability map that belong to credit unit g. We reduce these entries, apply the question-specific credit map, and form a single weighted policy ratio:

$$\begin{aligned} u_{i,t,g}^{\theta} &= \frac{1}{|\Omega_g|} \sum_{r \in \Omega_g} \log \pi_\theta(\mathbf{z}_{i,t-1}^{r} \mid \mathbf{z}_{i,t}, t, c), \\ \ell_{i,t,j}^{\theta} &= \sum_{g \in \mathcal{G}_\Gamma} w_{i,j,g} u_{i,t,g}^{\theta}, \qquad \rho_{i,t,j}(\theta) = \exp\Big(\ell_{i,t,j}^{\theta} - \ell_{i,t,j}^{\theta_{\mathrm{old}}}\Big)\,. \end{aligned} \tag{5}$$

Our token-level GRPO surrogate is

$$\mathcal{L}_{\mathrm{tvrl}} = \frac{1}{TMK} \sum_{i,t,j} \max\Big(-\hat{A}_i \rho_{i,t,j}(\theta), -\hat{A}_i \,\mathrm{clip}(\rho_{i,t,j}(\theta), 1-\epsilon, 1+\epsilon)\Big)\,, \tag{6}$$

where $\hat{A}_i = \mathrm{clip}(A_i^{\mathrm{tvrl}}, -A_{\max}, A_{\max})$; we omit the inherited timestep-balancing factors for readability. Log-probability changes are aggregated before exponentiation and clipping. With uniform final latent-grid weights and identical check maps, the weighted log-probability reduces to the scalar baseline's reduction. By contrast, exponentiating and clipping a separate ratio for every unit would change the estimator, and this variant performs worse in our ablation (Appendix A.4).



Figure 4: **Qualitative results on HunyuanVideo-1.5.** Comparison of prompt- and seed-matched videos from the base model (Wu et al., 2025a), Dance-GRPO (Xue et al., 2025), Flow-GRPO (Liu et al., 2025a), SAGE-GRPO (Zheng et al., 2026), and TVRL with the Qwen3.5-9B critic on SAGE-GRPO validation prompts (full text in Appendix A.2); row labels name the prompt requirement to check. Only TVRL satisfies all three, whereas the baselines substitute a flat zither, put a fist or an object in the raised hand, or lose the low-angle framing and the dress.

The weighted ratio is a local credit-assignment surrogate rather than the exact likelihood ratio of the full denoising trajectory. In the unclipped region, its gradient takes the form

$$\nabla_\theta \mathcal{L}_{i,t,j} \propto -\hat{A}_i \rho_{i,t,j}(\theta) \sum_{g \in \mathcal{G}_\Gamma} w_{i,j,g} \nabla_\theta u^\theta_{i,t,g}. \tag{7}$$

Hence $\hat{A}_i$ determines the sign and overall magnitude of the update, while $w_{i,j,g}$ routes a larger share of it to the units most sensitive to check j. At the old policy every $\rho_{i,t,j} = 1$, so to first order the K maps act as their average $\bar{w}_{i,g} = \frac{1}{K}\sum_j w_{i,j,g}$; they differ only through exponentiation and clipping, and a single ratio with $\bar{w}$ performs worse (Appendix A.4). Appendix B.7 gives conditional intuitions for when routing helps.

3.5 KL REGULARIZATION

As in SAGE-GRPO (Zheng et al., 2026), we add reference and stepwise old-policy KL terms:

$$\begin{aligned} \mathcal{D}_{\text{ref}} &= \mathbb{E}_{i,t}\left[\frac{\langle \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|_2^2 \rangle}{2\sigma_{\eta,t}^2}\right], \\ \mathcal{D}_{\text{step}} &= \mathbb{E}_{i,t}[\log \pi_{\theta_{\text{old}}}(\mathbf{z}_{i,t-1} \mid \mathbf{z}_{i,t}, t, c) - \log \pi_\theta(\mathbf{z}_{i,t-1} \mid \mathbf{z}_{i,t}, t, c)], \\ \mathcal{L} &= \mathcal{L}_{\text{tvrl}} + \beta_{\text{KL}}\left(\lambda_{\text{ref}}\mathcal{D}_{\text{ref}} + \lambda_{\text{step}}\mathcal{D}_{\text{step}}\right), \end{aligned} \tag{8}$$

where $\langle\cdot\rangle$ averages over non-batch entries and an adaptive controller updates β_{KL}. Both regularizers remain global over transition states; token credit affects only the clipped policy-ratio term. The coefficients are listed in Appendix B.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Training setup. We fine-tune HunyuanVideo-1.5 (Wu et al., 2025a) on the public SAGE-GRPO prompts (Zheng et al., 2026), keeping the VAE decoder, the text encoder, and the VLM critic frozen. Training rollouts contain 81 frames at 352×624 resolution with 20 denoising steps; evaluation videos contain 121 frames at 480×864 with 40 steps. We use the Muon optimizer with learning rate 10^{-5} and weight decay 0.01 on 64 NVIDIA A100 80GB GPUs, with a per-device batch size of 2 and a group size of 4. The default Qwen3.5-9B critic scores five offline-decomposed yes/no checks by teacher-forced answer log-probability. Their average forms the shared group-relative advantage, and their credit maps route it in the policy ratio.

Table 1: VBench-2.0 results with the SAGE sampler. Per reward model, GRPO uses uniform routing and TVRL gradient-based token routing; Δ is TVRL's Overall gain.

Reward	Method	Overall	Δ	Creat.	Comm.	Ctrl.	Human	Phys.
Base model	–	54.09	–	41.40	62.75	30.26	88.94	47.11
VideoAlign (Liu et al., 2025b)	GRPO	54.18	–	41.44	61.14	30.77	90.06	47.49
	TVRL	55.54	**+1.36**	45.11	61.16	32.09	90.21	49.15
VideoScore2 (He et al., 2025)	GRPO	54.66	–	42.23	64.89	30.29	91.52	44.35
	TVRL	55.99	**+1.33**	42.08	64.60	31.33	90.79	49.13
UnifiedReward2 (Wang et al., 2025b)	GRPO	54.82	–	42.90	62.14	30.25	88.90	50.90
	TVRL	56.67	**+1.85**	45.36	64.55	31.29	89.87	52.28
Qwen3.5-9B (Qwen Team, 2026)	GRPO	54.54	-	41.68	64.88	31.57	88.85	45.74
	TVRL	**57.69**	**+3.15**	47.36	64.31	31.64	90.76	54.37

Baselines and variants. Table 1 fixes the SAGE sampler and compares four reward models. GRPO uses each scalar reward with uniform routing, whereas TVRL additionally uses the video-frame-input gradient of the same reward for token routing. We also vary the SDE sampler and ablate credit granularity and the critic (more controls in Appendix A).

Evaluation. We report the VBench-2.0 Overall score together with its creativity, common sense, controllability, human action, and physics dimensions (Zheng et al., 2025). Our pairwise human study uses the first 200 prompts from VideoGen-Eval (Yang et al., 2025) and three distinct seeds per model and prompt, yielding 600 videos per model; eight annotators provide blind win, loss, or tie judgments of text–video alignment. Appendix A.1 describes the full protocol. All evaluated checkpoints are taken after 100 optimizer steps, and unless stated otherwise, table entries and optimization curves average three random seeds and std showed in Appendix A.3.

4.2 MAIN RESULTS

Across the tested configurations, TVRL improves over matched GRPO baselines with several stochastic samplers and differentiable evaluators. It raises VBench-2.0 Overall with the SAGE, Flow, and Dance samplers and yields positive gains for all four evaluators tested under SAGE. These results establish empirical compatibility in the evaluated settings; they do not imply sampler- or evaluator-independent guarantees.

SDE sampler compatibility. Figure 5 fixes Qwen3.5-9B as the reward and credit model and varies the SDE sampler. TVRL raises Overall from 54.54 to 57.69 with SAGE (+3.15) and from 53.79 to 56.71 with Flow (+2.92). With Dance, the completed GRPO runs average 50.84, compared with 53.52 for TVRL (+2.68). Because the evaluator and the routing rule are held fixed, these gains indicate compatibility with different stochastic transitions rather than a change in the reward.

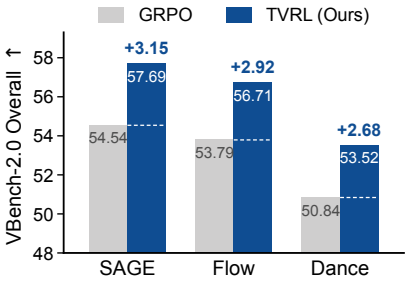

Figure 5: VBench-2.0 Overall of GRPO and TVRL across SDE samplers.

Reward model compatibility. Table 1 holds the SAGE sampler fixed. TVRL improves Overall from 54.18 to 55.54 (+1.36) with VideoAlign, from 54.66 to 55.99 (+1.33) with VideoScore2, from 54.82 to 56.67 (+1.85) with UnifiedReward2, and from 54.54 to 57.69 (+3.15) with Qwen3.5-9B. Each pair uses the same evaluator and differs only in whether its video-frame-gradient map routes the policy up-

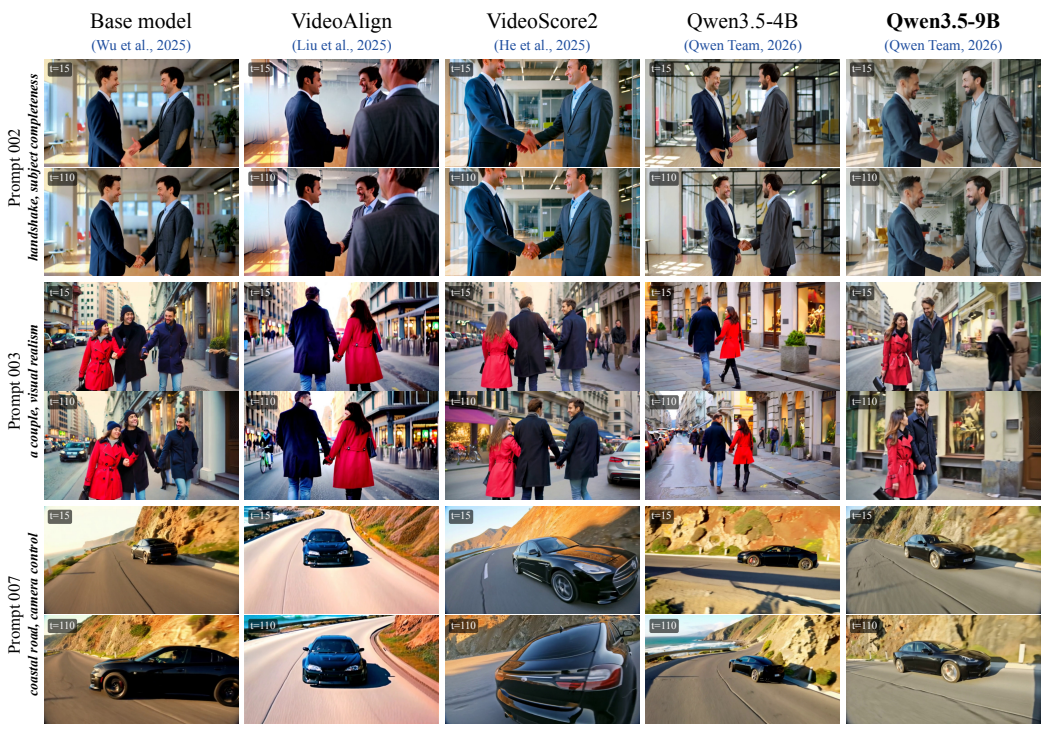

Figure 6: **Effect of the frozen critic.** Prompt- and seed-matched videos from the base model (Wu et al., 2025a) and TVRL with VideoAlign, VideoScore2, Qwen3.5-4B, or Qwen3.5-9B as the critic on VideoGen-Eval prompts (Yang et al., 2025) (Appendix A.2); row labels name the aspect to check. Only the Qwen3.5-9B critic keeps all three, whereas weaker critics add a third person, show the couple from behind, or lose the road.

date, so these three-seed means are consistent with a benefit from gradient-based routing. With Qwen3.5-9B the gain comes mainly from creativity (+5.68) and physics (+8.63), while controllability is unchanged (31.57 to 31.64) and common sense drops slightly (64.88 to 64.31), suggesting that the size of the gain depends both on the scalar objective and on whether its video sensitivity provides useful routing. More controls and results with other samplers are in Appendices A and B.4.

Human preference. Table 2 summarizes the blind pairwise judgments. Participants prefer TVRL more often than the matched GRPO baseline under SAGE (36.3% versus 23.3%) and more often than the base model (31.3% versus 19.2%). The remaining 40.4% and 49.5% of judgments are ties, indicating that many pairs are perceptually comparable. Excluding ties, TVRL wins 60.9% of the decisive comparisons against GRPO and 62.0% against the base model, consistent with VBench-2.0.

Table 2: Blind pairwise human preferences for TVRL against the matched GRPO baseline under SAGE and the base model (%).

Pair	Win	Loss	Tie
vs. GRPO (SAGE)	**36.3**	23.3	40.4
vs. Base	**31.3**	19.2	49.5

Qualitative results. Figure 1 compares TVRL with the base model on prompt- and seed-matched examples: it renders a grand piano rather than a generic keyboard, places the requested denim jacket on a model, and corrects the twisted legs of a dancer. Figure 4 compares against Dance-GRPO, Flow-GRPO, and SAGE-GRPO on three prompt- and seed-matched SAGE-GRPO validation prompts (full prompts in Appendix A.2). The gains concentrate on the requirements the prompts spell out: only TVRL renders an upright erhu with the bow in motion, the man's empty hand counting on his fingers, and the low-angle garden shot with a coherent dress and the woman reaching toward a rose, whereas the baselines distort the erhu, lose the gesture, or collapse the dress. These examples illustrate, but do not substitute for, the quantitative and human studies.

4.3 Ablation Study

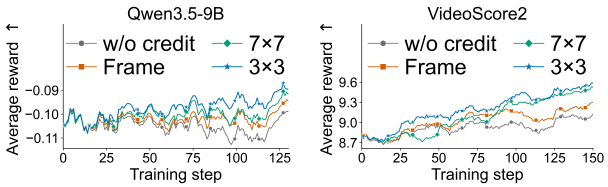

Figure 7: **Finer credit learns faster.** Training reward with Qwen3.5-9B (left) and VideoScore2 (right), averaged over three runs: w/o credit < frame < 7×7 < 3×3 later in training.

Credit granularity. Figure 7 compares progressively localized credit: uniform weighting (w/o credit), frame-level credit, and 7×7 and 3×3 spatial credit. Averaged over three runs, the Qwen3.5-9B and VideoScore2 rewards both follow w/o credit < frame < 7×7 < 3×3 in later training, which supports progressively localized routing within the tested range. It does not imply that finer credit is always better: the unsmoothed 1×1 control reaches 55.10 Overall with Qwen3.5-9B, above uniform routing (54.54) but below frame-level credit (56.47; Appendix B.8). Table 3 provides complementary held-out evidence: Overall rises from 54.54 with uniform routing to 56.47 with frame credit, 57.08 with 7×7, and 57.69 with 3×3. Shuffling the 3×3 weights reduces Overall to 53.70, showing that the locations and coherence of the weights matter; since shuffling precedes smoothing and alters the final distribution, it does not isolate semantic alignment. We further use DINOv2 features as a reward-agnostic source of the routing map (Ni et al., 2026), keeping all other settings identical to our 3×3 setting. The result is 55.48 (DINOv2 map): nonuniform routing helps even without the reward, and deriving the map from the reward helps further.

Frozen VLM critic. Table 4 tests sensitivity to the frozen VLM used jointly for reward and credit. Within the Qwen3.5 family, Overall rises monotonically from 51.85 with the 0.8B model to 56.02 with 4B and 57.69 with 9B. Critics from other families show a similar ordering: InternVL3-1B obtains 53.98 and Gemma4-E4B-it reaches 55.98. Because changing the critic changes both its scalar reward and its gradient map, this experiment cannot attribute the gain to either component alone; it shows that the complete TVRL signal benefits from the larger critics tested, not a general scaling law. Figure 6 shows the same trend qualitatively.

Table 3: Routing ablation (SAGE, Qwen3.5-9B).

Routing	Overall	Δ
Uniform	54.54	–
Frame	56.47	+1.93
7×7	57.08	+2.54
3×3 shuffled	53.70	-0.84
DINOv2 map	55.48	+0.94
3×3 **(Ours)**	**57.69**	**+3.15**

Table 4: Frozen VLM critic (SAGE).

Critic	Overall	Δ
Qwen3.5-0.8B	51.85	-5.84
InternVL3-1B	53.98	-3.71
Gemma4-E4B-it	55.98	-1.71
Qwen3.5-4B	56.02	-1.67
Qwen3.5-9B	**57.69**	–

Spatial credit diagnostics. Figure 8 shows a qualitative example for Table 3 of why 3×3 credit is used by default. The 1×1 map is noisy: its residual over 3×3 (left) overemphasizes the background tokens unrelated to the check, consistent with the variance argument in Appendix B.7. The 3×3 map is sharper than 7×7: its residual (right) lies near the key of the question, the sunglasses.

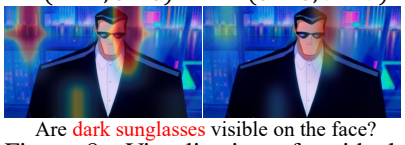

Figure 8: Visualization of residuals $\Delta(1 \times 1, 3 \times 3)$ and $\Delta(3 \times 3, 7 \times 7)$.

5 Conclusion

We presented TVRL, a framework that derives token-level credit from the reward being optimized in video reinforcement learning, and instantiated it in GRPO for text-to-video diffusion. Averaged teacher-forced question rewards set the sign and strength of each update through a shared group-relative advantage, and detached VLM gradients set its spatio-temporal allocation inside the clipped policy ratio. On VBench-2.0, TVRL improves over matched GRPO baselines for all four reward models under SAGE (+1.33 to +3.15 Overall) and for all three samplers with Qwen3.5-9B (+2.68 to +3.15), and 3×3 spatial credit outperforms both coarser and unsmoothed 1×1 credit. TVRL shows a promising direction for video RL: moving from trajectory-level to token-level updates, so that each rollout's advantage acts on the video tokens that determine its reward. We expect token-level credit to extend to longer videos and other evaluators.

AI USE STATEMENT

In this work, we used generative AI tools (large language model assistants) to help implement and debug code and to assist with analysis scripts that aggregate the reported results. We did not use generative AI tools to propose or refine the research hypotheses, to design the method or the experiments, to formulate mathematical claims or to write proofs, to generate synthetic datasets, or to interpret the results. The remaining tasks with required disclosure (assisting with translation, cleaning or reformatting datasets, and supporting qualitative or thematic data analysis) are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript for readability, to format LaTeX tables and figure code, and to check and format bibliographic references. We note that the frozen vision–language critics used as reward and credit models (Section 3) are components of the proposed method rather than authoring aids, and their role is fully described in the paper. We have reviewed all AI-assisted work: AI-assisted code was tested by the authors before use, AI-edited text was checked by the authors against the underlying experimental records, and every reference was verified against its published source. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work includes a human preference study (Section 4.1 and Appendix A.1) in which eight annotators compared pairs of generated videos for text–video alignment. Annotators viewed only synthetic videos generated from publicly available benchmark prompts, model identities were hidden, and no personal or sensitive data were collected from or about the annotators. All base models, reward models, prompt sets, and benchmarks used in this work are publicly released by their respective authors, and we do not release any new dataset. TVRL improves the prompt adherence of an existing text-to-video generator and therefore inherits the general risks of video generation, including the potential to produce misleading or inappropriate content; it does not add new generative capabilities beyond those of the underlying base model and evaluators. Because the VLM critic is used as a frozen source of rewards and token credit, any biases in its judgments may be transferred into the fine-tuned generator; we regard this as a limitation shared with all reward-driven post-training and recommend auditing the critic before deployment. The authors declare no conflicts of interest beyond their institutional affiliations.

REPRODUCIBILITY STATEMENT

We will release code and checkpoints publicly. Section 4.1 specifies the base model, training prompts, rollout and evaluation resolutions, optimizer, hardware, batch and group sizes, and the default critic. Appendix B.2 lists the complete hyperparameter configuration of the main run, and Appendix B.1 gives the SDE transition and the log-probability used in the policy ratio. The exact prompt-decomposition template and the VLM reward prompt with its reference answer are reproduced verbatim in Appendices B.5 and B.6, and the derivation of the token-credit objective is given in Appendix B.7. The automatic and human evaluation protocols are described in Appendices A.1 and B.3. All base models, reward models, prompt sets, and benchmarks are publicly available, and all reported numbers are means over three random seeds unless stated otherwise.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. Videoscore2: Think before you score in generative video evaluation, 2025. URL https://arxiv.org/abs/2509.22799.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *CVPR*, 2024.

Yuanzhi Liang, Xuan'er Wu, Yirui Liu, Yijie Fang, Yizhen Fan, Ke Hao, Rui Li, Ruiying Liu, Ziqi Ni, Peng Yu, Yanbo Wang, Haibin Huang, Qizhen Weng, Chi Zhang, and Xuelong Li. Teleboost: A systematic alignment framework for high-fidelity, controllable, and robust video generation, 2026. URL https://arxiv.org/abs/2602.07595.

Lightricks. LTX-2: Open-source ai video and audio generation. https://www.lightricks.com/ltxv, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

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

Ziqi Ni, Yuanzhi Liang, Rui Li, Yi Zhou, Haibin Huang, Chi Zhang, and Xuelong Li. Seeing what matters: Visual preference policy optimization for visual generation, 2026. URL https://arxiv.org/abs/2511.18719.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

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

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

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.

Shuting Wang, Haihong Tang, Zhicheng Dou, and Chenyan Xiong. Harness local rewards for global benefits: Effective text-to-video generation alignment with patch-level reward models, 2025a. URL https://arxiv.org/abs/2502.06812.

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

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, 2026. URL https://arxiv.org/abs/2601.04153.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. Hunyuanvideo 1.5 technical report, 2025a. URL https://arxiv.org/abs/2511.18870.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025b.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. *NeurIPS*, 2024.

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. *NeurIPS*, 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, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

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.

Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. Videogen-eval: Agent-based system for video generation evaluation. *arXiv preprint arXiv:2503.23452*, 2025.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. InstructVideo: Instructing video diffusion models with human feedback. In *CVPR*, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, 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.

Mingzhe Zheng, Weijie Kong, Yue Wu, Dengyang Jiang, Yue Ma, Xuanhua He, Bin Lin, Kaixiong Gong, Zhao Zhong, Liefeng Bo, Qifeng Chen, and Harry Yang. Manifold-aware exploration for reinforcement learning in video generation, 2026. URL https://arxiv.org/abs/2603.21872.

Supplementary Material for Token-Level Video Reinforcement Learning

A Additional Experimental Results

A.1 Human evaluation protocol

We use the first 200 prompts of VideoGen-Eval (Yang et al., 2025). For every prompt, each compared model generates three videos from three distinct seeds, yielding 600 videos per model. The two pairwise comparisons evaluate TVRL against the matched GRPO baseline under SAGE and against the pretrained base model. Eight annotators complete the blind study with model identities hidden and judge text–video alignment as a win for TVRL, a loss, or a tie. We aggregate all judgments into the win/loss/tie rates in Table 2; the decisive preference rate excludes ties and is computed as $\mathrm{Win}/(\mathrm{Win}+\mathrm{Loss})$.

A.2 Prompts for the qualitative comparison

The three prompts in Figure 4 are validation prompts 005, 015, and 016 of the public SAGE-GRPO prompt set (Zheng et al., 2026), and the three prompts in Figure 6 are prompts 002, 003, and 007 of VideoGen-Eval (Yang et al., 2025). Within each figure, all methods generate from the same prompt and seed. The full prompt text follows.

Prompt 005 (SAGE-GRPO validation).

```
The scene opens with a medium shot of an elderly woman, appearing to
    be in her 80s, seated on a simple wooden stool in a public park.
    She has fine, silver hair neatly tied back and wears a traditional
    , dark-colored tunic with intricate embroidery at the collar. Her
    eyes are gently closed as she focuses on her music. In her hands,
    she holds a polished, dark wood erhu. Her left hand delicately
    presses the strings, her fingers moving with practiced grace,
    while her right hand draws the bow smoothly across them, producing
     a soulful, resonant melody. She hums softly, a low, gentle sound
    that harmonizes perfectly with the instrument's voice. A diverse
    crowd of onlookers, ranging from young adults to other seniors,
    has formed a respectful semi-circle around her. Some watch with
    rapt attention, their expressions contemplative. A young woman can
     be seen subtly recording on her phone, while an older man nods
    his head slowly to the rhythm. The park setting is serene, with
    large, mature trees whose leaves create a soft, dappled canopy.
    The light is diffused, characteristic of an overcast day, which
    lends a moody and introspective quality to the entire scene. The
    focus remains tightly on the woman, a figure of serene artistry
    amidst the quiet reverence of her audience.
```

Prompt 015 (SAGE-GRPO validation).

```
The scene opens on a sweeping wide shot, composed according to the
    rule of thirds, with an elderly Chinese man positioned on the
    right side of the frame. He stands on an ancient, weathered stone
    path that meanders through a landscape of rolling hills shrouded
    in a thick, soft mist. The man has a deeply wrinkled face and a
    long, flowing white beard that reaches his chest. He wears a
    simple, coarse-spun brown robe and a traditional conical bamboo
    hat that casts a gentle shadow over his features. In his left hand
```

```
    , he grips a gnarled wooden cane, its surface polished smooth from
     years of use, providing him with stability. His entire focus is
    on his right hand, which is held up near his face. His brow is
    furrowed in concentration as his thumb methodically touches
    different segments of his fingers in a complex, silent calculation
    . The environment is lit by a diffused, natural light coming from
    the side, which catches the airborne mist and softly illuminates
    the contours of his robe and the sparse, gnarled trees dotting the
     background. The atmosphere is one of profound peace and deep
    contemplation, as if the man is solving a timeless riddle alone in
     nature.
```

Prompt 016 (SAGE-GRPO validation).

```
The scene opens with a low-angle, full shot of a woman with dark,
    flowing hair, wearing a striking, floor-length scarlet red dress.
    She is strolling leisurely from right to left along a winding
    cobblestone path in a magnificent, verdant garden. The garden is
    in full bloom, with an array of colorful flowers like deep pink
    roses and tall, yellow sunflowers flanking the path. The camera is
     positioned low to the ground, making the flowers in the
    foreground appear large and immersive, framing the woman
    majestically. It's the golden hour; the setting sun is out of
    frame but casts long shadows and bathes the scene in a warm,
    golden light. This backlights the woman, creating a soft, ethereal
     glow around her silhouette and illuminating the delicate fabric
    of her dress. As she walks, she gently brushes her fingertips
    against the petals of a rose. The camera pans smoothly, following
    her elegant and unhurried pace. The background consists of tall
    green hedges and a soft-focus view of the sky, painted with hues
    of orange and pink.
```

Prompt 002 (VideoGen-Eval).

```
In a sleek, modern office with glass walls and minimalist decor, two
    men stand facing each other, their expressions radiating warmth
    and camaraderie. One man, wearing a tailored navy suit with a
    crisp white shirt, extends his hand with a genuine smile. The
    other, dressed in a charcoal blazer and light blue shirt,
    reciprocates the gesture, their hands meeting in a firm,
    enthusiastic handshake. Sunlight streams through large windows,
    casting a soft glow on the polished wooden floor and contemporary
    furniture. The atmosphere is one of mutual respect and shared
    success, captured in this moment of professional connection.
```

Prompt 003 (VideoGen-Eval).

```
A couple strolls hand in hand down a bustling city street, their
    fingers intertwined, embodying a sense of unity and affection. The
     man, wearing a navy coat and jeans, walks confidently beside the
    woman, who is dressed in a stylish red trench coat and black boots
    . The street is alive with the sounds of distant traffic and the
    chatter of passersby, while the couple remains in their own world,
     exchanging smiles and glances. The static camera captures their
    every step, as they navigate the urban landscape, framed by
    towering buildings and vibrant storefronts, under a sky painted
    with the warm hues of a setting sun.
```

Prompt 007 (VideoGen-Eval).

```
A sleek black car glides smoothly along a winding coastal road, the
    camera capturing its every move with a handheld tracking shot. The
     sun casts a golden glow on the vehicle's polished surface,
    highlighting its aerodynamic design. As the car navigates sharp
    turns, the ocean waves crash against the rocky cliffs below,
    creating a dramatic backdrop. The camera follows closely,
    capturing the rhythmic motion of the wheels and the subtle
    reflections of the surrounding landscape on the car's glossy
    exterior. The scene conveys a sense of freedom and adventure, with
     the open road stretching endlessly ahead.
```

A.3 SEED STATISTICS

Tables A1 and A2 report the mean ± standard deviation behind the main-text numbers, over n runs that use independent training seeds and independent generation seeds, each evaluated at step 100 on the same prompt set.

Table A1: Table 1 with standard deviations over seeds (SAGE sampler, step 100).

Reward	Method	Overall	Creat.	Comm.	Ctrl.	Human	Phys.	n
Base model	–	54.09 ±0.11	41.40 ±0.38	62.75 ±0.24	30.26 ±0.46	88.94 ±0.13	47.11 ±0.34	3
VideoAlign	GRPO	54.18 ±0.23	41.44 ±0.41	61.14 ±0.26	30.77 ±0.44	90.06 ±0.12	47.49 ±0.35	3
	TVRL	55.54 ±0.42	45.11 ±0.36	61.16 ±0.27	32.09 ±0.43	90.21 ±0.14	49.15 ±0.32	3
VideoScore2	GRPO	54.66 ±0.16	42.23 ±0.39	64.89 ±0.22	30.29 ±0.48	91.52 ±0.11	44.35 ±0.37	3
	TVRL	55.99 ±0.33	42.08 ±0.40	64.60 ±0.21	31.33 ±0.45	90.79 ±0.13	49.13 ±0.31	3
UnifiedReward2	GRPO	54.82 ±0.47	42.90 ±0.37	62.14 ±0.25	30.25 ±0.49	88.90 ±0.15	50.90 ±0.30	3
	TVRL	56.67 ±0.19	45.36 ±0.34	64.55 ±0.23	31.29 ±0.46	89.87 ±0.12	52.28 ±0.28	3
Qwen3.5-9B	GRPO	54.54 ±0.38	41.68 ±0.42	64.88 ±0.20	31.57 ±0.44	88.85 ±0.16	45.74 ±0.36	3
	TVRL	57.69 ±0.13	47.36 ±0.33	64.31 ±0.22	31.64 ±0.47	90.76 ±0.11	54.37 ±0.27	3

Table A2: Overall mean ± standard deviation for the sampler comparison (Figure 5), the routing ablation (Table 3), and the critic ablation (Table 4), all with the Qwen3.5-9B reward. Dance GRPO averages two runs; all other rows average three.

	Setting	Overall
Sampler	SAGE, GRPO	54.54 ±0.38
	SAGE, TVRL	57.69 ±0.13
	Flow, GRPO	53.79 ±0.31
	Flow, TVRL	56.71 ±0.27
	Dance, GRPO	50.84 ±0.42
	Dance, TVRL	53.52 ±0.34
Routing	Uniform	54.54 ±0.38
	Frame	56.47 ±0.29
	7×7	57.08 ±0.20
	3×3	57.69 ±0.13
	3×3 shuffled	53.70 ±0.46
	DINOv2 map	55.48 ±0.32
Critic	Qwen3.5-0.8B	51.85 ±0.41
	InternVL3-1B	53.98 ±0.37
	Gemma4-E4B-it	55.98 ±0.28
	Qwen3.5-4B	56.02 ±0.24
	Qwen3.5-9B	57.69 ±0.13

A.4 ADDITIONAL ABLATIONS

Reward objective and token-level credit. Table A3 compares several SAGE/Qwen3.5-9B reward-objective variants that precede the final shared-advantage setting. Free-generation scoring, target log-probability scoring, and token credit obtain similar Overall scores, ranging from 54.28

Table A3: Reward-objective ablations preceding the final shared-advantage setting. All rows use SAGE-GRPO with Qwen3.5-9B and the same decomposed binary-question prompts, while varying VLM scoring mode and temporal credit assignment.

Variant	Reward	Credit	Overall	Creat.	Comm.	Ctrl.	Human	Phys.
Free generation	`Free-gen`	Uniform	54.28	37.32	67.75	29.75	87.42	49.14
Log-probability reward	`Logprob`	Uniform	54.54	41.68	64.88	31.57	88.85	45.74
Token credit	`Logprob`	Per-question gradient	54.42	39.98	66.31	30.38	88.41	47.02

Table A4: Rating-threshold, reward-frame-count, and question-count ablations for SAGE-GRPO with Qwen3.5-9B token credit, evaluated after 100 optimization steps. All score columns report VBench-2.0 results.

Control	Setting	Overall	Creat.	Comm.	Ctrl.	Human	Phys.
Rating threshold	5	55.41	36.98	69.20	33.03	88.71	49.14
Reward frames	5	53.49	40.81	65.47	30.42	87.08	43.64
	30	55.71	44.60	61.37	30.41	90.02	52.14
Questions	1	51.34	33.52	64.34	27.91	87.62	43.30
	3	55.69	40.24	65.43	31.58	90.20	51.00

to 54.54. The teacher-forced reward alone therefore does not explain the main-result gain. The stronger result in Table 1 comes from combining VLM answer likelihoods with the shared rollout-level advantage and weighted-log-probability token credit.

Reward-frame and question controls. Table A4 examines how much video and semantic evidence the critic uses. Increasing the number of sampled reward frames from 5 to 30 improves Overall from 53.49 to 55.71, with the largest gains in creativity, human action, and physics. Increasing the number of decomposed questions from 1 to 3 produces an even larger Overall gain, from 51.34 to 55.69, suggesting that broader semantic coverage matters for both reward estimation and credit assignment. The rating-threshold-5 variant reaches 55.41 Overall; because only one threshold is reported, we treat this as a sensitivity result rather than evidence for an optimal threshold.

Gradient aggregation and ratio form. Table A5 compares alternatives at frame-level credit, where per-question routing reaches 56.47. Because every ratio equals one at the old policy, per-question maps and their average $\bar{w}$ give the same first-order update; the averaged-map single ratio isolates what separate ratios add through exponentiation and clipping, and it drops to 54.96. A generic check in place of the prompt-specific ones also reaches 54.96, and clipping a separate ratio for each frame drops to 52.95, supporting a single weighted ratio per check (Section 3.4).

A.5 Training overhead

TVRL incurs approximately 18% higher observed time per update in our historical SAGE runs with VideoAlign (Table A6); this pair does not measure the cost of the default Qwen3.5-9B configuration. We compare the scalar-reward GRPO baseline with the six-frame VideoAlign credit implementation using the shared rollout-level advantage and weighted-log-probability routing. Both runs use seed 42, batch size 2 per device, group size 4, sequence-parallel size 2, and 20 denoising steps, with the same logged optimizer and KL settings. We compute elapsed time from consecutive training-record timestamps over updates 11–100, excluding the first ten updates and retaining all 90 subsequent intervals. This measurement includes work outside the rollout and optimizer-loop timers, as well as any logging or checkpoint activity within these intervals.

The observed mean increases from 666.05 to 786.04 seconds per update, a difference of 119.99 seconds (18.02%). This is an implementation-level wall-clock comparison, not an isolated measurement of VLM backward cost. In particular, the credit implementation also recomputes old-policy log-probabilities for weighted routing; this required work is included in the reported elapsed time.

Table A5: Ablations on how VLM-gradient credit is constructed and applied, for SAGE-GRPO with Qwen3.5-9B evaluated after 100 optimization steps. All rows use frame-level credit; per-question frame credit reaches 56.47. General semantic credit uses one generic check; averaged map uses one ratio with $\bar{w} = \frac{1}{K}\sum_j w_j$, the first-order equivalent of per-question ratios; per-frame ratio clips frame-wise ratios before aggregation.

Ablation	Overall	Creat.	Comm.	Ctrl.	Human	Phys.
General semantic credit	54.96	44.00	66.31	27.41	88.46	48.64
Averaged map, single ratio	54.96	42.40	64.88	34.71	86.94	45.88
Per-frame ratio	52.95	38.77	66.03	29.04	84.66	46.26

Table A6: Observed training overhead in historical SAGE/VideoAlign runs. Times are seconds per update over updates 11–100; relative time is the ratio of means. Configuration differences are described in Appendix A.5.

Method	Mean (s)	Median (s)	Relative time
GRPO	666.05	661.72	1.000×
TVRL	786.04	781.53	1.180×

The comparison uses existing runs rather than a controlled timing rerun: the baseline and credit runs use VideoAlign MQ/TA/VQ reward weights of $(0.5, 1, 0.5)$ and $(1, 1, 1)$, respectively, and different prompt-file variants (original versus question-decomposed). The baseline reward-frame count is inherited from its reward-model defaults, while the credit run explicitly uses six frames; identical historical frame sampling, source revisions, and execution conditions have not been verified. We therefore report 18.02% as the observed overhead for this pair, rather than a universal overhead estimate or evidence of a wall-clock speedup.

A.6 Question-conditioned spatial sensitivity

Relation to the training control. The shuffled 3×3 control in Table 3 obtains 53.70 Overall, compared with 57.69 for unshuffled routing and 54.54 for uniform routing. In the local implementation, shuffling permutes routing-grid weights before spatial smoothing, preserving their value distribution at that stage; subsequent smoothing can change the distribution. This comparison supports structured routing but does not isolate semantic correspondence from spatial and temporal coherence or match the final maps' entropy. We separately test whether the critic's video sensitivity follows queried content when its position changes.

Controlled question and position swaps. We fix 12 disjoint pairs of source prompts and two generation seeds per pair, yielding 24 video pairs. Each source contributes 20 uniformly sampled frames, center-cropped to a square and resized to fill an equal-size panel; the crop is fixed and not object-aware. An earlier construction instead letterboxed each full frame into its panel with neutral padding, which left the content in roughly half of the panel area; we report it as a second row on the same pairs and questions. For each pair, we concatenate the panels as AB and BA and ask separately whether target A or target B is present. The questions do not specify positions. This gives 96 real Qwen3.5-9B gradient computations per panel construction of the teacher-forced log-probability of the fixed answer `Yes`. Unlike the full training question set, this diagnostic uses object-presence checks. Before computing gradients, we review three frames per pair and require each target to be visible in at least two. This qualifies 20 crop-to-fill pairs: one pair contains an ambiguous piano-like object, and in three pairs the fixed crop removes the target from at least two reviewed frames. The letterbox construction qualifies 23 pairs under the same rule. Unqualified pairs are retained in the all-pairs analysis, every prompt-pair group keeps at least one qualified pair, and no pairs are replaced after observing gradients. The construction uses known panel boundaries rather than detector or segmenter masks; each panel nevertheless contains background as well as the target.

Coordinate reconstruction and endpoint. The diagnostic uses the training critic configuration: 20 frames under a 50,176-pixel budget shared across the clip. The actual critic input is therefore $20 \times 32 \times 64$, producing a $10 \times 2 \times 4$ gradient grid in which each panel spans only 2×2 patches per temporal slot. We invert the processor's block-major patch packing to recover spatial coordinates, then apply the training helpers' interpolation to $20 \times 16 \times 16$ and 3×3 smoothing. Interpolation and smoothing leak credit across the panel boundary: a map placing all native-grid mass in one panel yields $D = 84.8$ pp, not 100, so the endpoint is mildly attenuated but far from saturated. Let $L(q, AB)$ denote normalized credit mass in the left panel under question q and layout AB. Define

$$D = \tfrac{1}{2}\big[L(A, AB) - L(B, AB) + L(B, BA) - L(A, BA)\big]. \tag{A1}$$

Positive D indicates that changing the question favors the corresponding panel, averaged over both positions. A score-independent saliency or foreground map gives zero within-layout differences; a fixed question-specific side preference cancels across layouts. Uniform routing has $D = 0$, and mean queried-panel mass is $1/2 + D/2$. We average seeds within each prompt-pair group and weight the 12 groups equally. Confidence intervals use 10,000 group bootstrap resamples; two-sided sign-flip tests operate on group means, not individual frames or gradient calls.

Table A7: Question–position crossover diagnostic with Qwen3.5-9B at the training critic resolution. D measures question-induced panel-credit change in percentage points (pp); the final column is mean credit mass in the queried panel, which equals 50% under uniform routing. Every cohort contains 12 prompt-pair groups. Crop-to-fill is the primary panel construction; letterbox is the earlier construction on the same pairs and questions.

Panels	Cohort	Pairs	D (pp), 95% CI	Queried mass
Crop-to-fill	Prequalified	20	$+5.39$ $[2.83, 7.77]$	52.70%
Crop-to-fill	All pairs	24	$+6.41$ $[4.46, 8.43]$	53.20%
Letterbox	Prequalified	23	$+3.79$ $[1.66, 5.99]$	51.90%
Letterbox	All pairs	24	$+3.74$ $[1.63, 5.90]$	51.87%

Results and scope. The qualified-cohort interaction is $+5.39$ percentage points (pp), with a 95% interval of $[2.83, 7.77]$ and sign-flip $p = 0.0034$. Mean queried-panel mass is 52.70%, only 2.70 pp above uniform. Ten of twelve groups have positive interactions, and both layout-specific contrasts are positive in 15 of 20 pairs. Including all 24 pairs gives $+6.41$ pp $[4.46, 8.43]$ with all twelve groups positive ($p = 0.0005$), and the letterbox construction gives a smaller effect of the same sign, $+3.79$ pp $[1.66, 5.99]$ (Table A7). The interaction is thus modest in magnitude but consistent in sign across groups and across both panel constructions, which supports question-conditioned regional sensitivity in this constructed setting; a score-independent map would give zero.

An exploratory decomposition, computed after the primary outcome was observed, indicates where the question dependence lies at this resolution. Within a fixed video, the two question-conditioned routed maps differ by a total variation of 30.5 pp on average. Their temporal marginals alone differ by 28.8 pp, whereas their spatial marginals differ by 8.3 pp, of which D is the component directed at the queried panel. With a 2×4 spatial grid per temporal slot, changing the question therefore mainly reallocates credit across frames and only secondarily across space. This agrees with the training ablation in Table 3, where frame credit accounts for most of the gain over uniform routing (54.54 to 56.47 Overall) and 3×3 spatial credit adds a smaller increment (to 57.69).

Because each panel contains background as well as the target, the result does not establish precise object segmentation or causal credit for generator updates. The deliberately selected split-screen examples and the low critic resolution limit generalization; the bootstrap assumes exchangeable groups and the sign-flip test assumes sign symmetry under the null. This diagnostic neither compares gradients with attention nor establishes that 3×3 is superior to 7×7.

B Additional Implementation Details

B.1 Diffusion policy details

Let $\mathbf{z}_T \sim \mathcal{N}(0, I)$. For a flow-matching velocity model, the deterministic Euler mean is

$$\boldsymbol{\mu}_\theta(\mathbf{z}_t, t, c) = \mathbf{z}_t + \Delta t_t \, \mathbf{v}_\theta(\mathbf{z}_t, t, c). \tag{B1}$$

During GRPO rollouts, the scheduler wraps this transition in a stochastic SDE step,

$$\mathbf{z}_{t-1} \sim \mathcal{N}\big(\boldsymbol{\mu}_\theta^{\mathrm{sde}}(\mathbf{z}_t, t, c), \sigma_{\eta,t}^2 I\big)\,, \tag{B2}$$

where η controls the injected sampling noise. Up to terms independent of θ, its transition log-probability is

$$\log \pi_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, t, c) = -\frac{\|\mathbf{z}_{t-1} - \boldsymbol{\mu}_\theta^{\mathrm{sde}}(\mathbf{z}_t, t, c)\|_2^2}{2\sigma_{\eta,t}^2} + \mathrm{const.} \tag{B3}$$

Different SDE samplers change the transition mean and the noise schedule but not the token-credit construction. We retain the unreduced elementwise log-probability map and aggregate it over the credit units as in Equation (5). The KL-regularized training objective is defined in Equation (8).

B.2 Training and reward settings

Table B1 summarizes the main TVRL setting used for the SAGE/Qwen3.5-9B token-credit video-adv run. Machine-specific paths, logging credentials, and storage locations are omitted. The Flow and Dance variants keep the same token-credit reward and GRPO configuration and change only the stochastic denoising sampler.

Table B1: Main training and reward settings.

Component	Setting
Generator	Text-to-video diffusion transformer, pretrained `480p_t2v` checkpoint
Post-training objective	GRPO with TVRL token credit
Main SDE type	`sage_grpo`
Reward critic	Frozen Qwen3.5-9B VLM
Reward model family	`qwen3_5`
Reward score type	`token_credit`
Reward VLM resolution budget	50,176 max pixels
Reward VLM batch size	1 generated video
Reward VLM generation cap	128 new tokens
Prompt checks	Five binary yes/no video-grounded checks per prompt
Reward frames	20 uniformly sampled frames per generated video
Group size	4 rollouts per prompt
Training rollout	81 frames at 352×624, 20 denoising steps
Evaluation generation	121 frames at 480×864, 40 denoising steps
Hardware	64 NVIDIA A100 80GB GPUs
Per-device batch size	2
Learning rate	1×10^{-5}
Optimizer	Muon optimizer, weight decay 0.01
LR schedule	Constant
Training budget	100 optimizer steps for every evaluated checkpoint
Timestep shift	5.0
Timestep permutation	Enabled
Gradient balancing	Enabled
Sequence parallel size	2
Memory options	FSDP and gradient checkpointing enabled
Loss mode	`video_adv`
Ratio mode	`weighted_logprob`
KL regularization	Moving reference KL and stepwise dual KL enabled
KL coefficients	KL weight 10^{-5}, initial adaptive coefficient 10^{-7}
Dual KL weights	Reference-moving weight 1.0, old-policy step weight 0.1
Reference update interval	10 optimizer steps
Checkpoint interval	25 optimizer steps in the current launcher

The training prompt file contains 10,000 text prompts. Each prompt is decomposed offline into an `overall_question`, five `sub_questions`, one structured VLM question prompt, and a structured reference answer. We do not use supervised target videos for post-training. During rollout, the generator samples videos for each text prompt, the frozen VLM scores the generated videos by teacher-forced answer likelihood, and the gradients of the same likelihoods provide the token-level credit weights.

B.3 Evaluation protocol

All runs are evaluated with the VBench-2.0 automatic metrics on the same held-out prompt split for matched baselines and TVRL variants. Tables list Overall first, followed by creativity, common sense, controllability, human action, and physics.

B.4 Supplementary VideoAlign sampler results

Table B2 repeats the sampler comparison of Figure 5 with VideoAlign in place of Qwen3.5-9B as the reward model. TVRL improves over the matched GRPO run under all three samplers: from 54.18 to 55.54 with SAGE (+1.36), from 55.07 to 55.82 with Flow (+0.75), and from 48.02 to 50.30 with Dance (+2.28). The gains are smaller than with Qwen3.5-9B (+3.15, +2.92, and +2.68), consistent with the main-text observation that the size of the improvement depends on the reward model whose gradient supplies the routing signal.

Table B2: VBench-2.0 Overall with VideoAlign as the reward model across SDE samplers; Δ is the gain of TVRL over GRPO.

SDE	Method	Overall	Δ
SAGE	GRPO	54.18	–
	TVRL	55.54	+1.36
Flow	GRPO	55.07	–
	TVRL	55.82	+0.75
Dance	GRPO	48.02	–
	TVRL	50.30	+2.28

B.5 Prompt decomposition template

We use an offline language model to convert each text-to-video prompt into five atomic yes/no checks. The decomposition model receives only the text prompt, not a generated video. The system message is:

System message

```
You are a structured data generator for text-to-video evaluation.
    Return exactly one valid JSON
object and nothing else.
```

The user template is:

Prompt decomposition template

```
You are an expert prompt-to-subquestion decomposition engine for text-
    to-video generation.

You are given a single text-to-video prompt, not a video and not an
    original evaluation question.
Your job is to convert the prompt itself into a compact binary
    checklist that can later be used to
verify whether a generated video faithfully follows the prompt.

Goal:
- Produce exactly <|NUM_SUBQUESTIONS|> atomic yes/no sub-questions.
- Each sub-question must target one concrete claim that is directly
    verifiable from the generated video and explicitly stated in
  the prompt.
- The sub-questions should together cover the main subject, scene,
    action or motion, key objects or
  attributes, and camera or composition when present.

Instructions:
1. Read the prompt carefully.
2. Rewrite the prompt intent as one concise 'overall_question'.
3. Produce exactly <|NUM_SUBQUESTIONS|> sub-questions.
4. Every sub-question must be:
```

```
   - directly verifiable from a generated video,
   - directly grounded in the prompt text,
   - phrased as a clear yes/no question,
   - atomic and non-redundant.
5. Prefer concrete video-observable facts:
   - subject identity or category,
   - visible objects,
   - explicit colors or materials,
   - spatial relations,
   - explicit actions or motion,
   - explicit camera motion or framing,
   - visible environment or background.
6. Do not ask about hidden intent, symbolism, or facts that cannot be
    observed in the video.
7. Avoid subjective or weakly grounded wording such as:
   - beautiful
   - impressive
   - dramatic
   - immersive
   - majestic
   - lonely
   - surreal
   - dreamlike
   unless the prompt gives an explicit manifestation that can be
    checked directly in the video.
8. Do not ask multiple things in one sub-question.
9. Do not repeat the same fact using slightly different wording.
10. If the prompt contains camera language, at least one sub-question
    should cover it.
11. If the prompt does not contain camera language, replace that slot
    with another explicit video-observable
    fact from the prompt.
12. Every `answer_type` must be `yes_no`.
13. Do not output answers.
14. Never return fewer than <|NUM_SUBQUESTIONS|> sub-questions.
15. If a prompt clause is broad, split it into smaller atomic video-
    grounded checks so that the final output
    still contains exactly <|NUM_SUBQUESTIONS|> items.
16. Prefer distributing the <|NUM_SUBQUESTIONS|> slots across subject,
     appearance, action,
    environment, and camera/composition when those aspects are present
    .

Output format:
- Return valid JSON only.
- No markdown.
- No code fences.
- No explanation outside the JSON.

Return JSON with this exact schema:
{
  "overall_question": "...",
  "sub_questions": [
    {
      "id": "q1",
      "question": "...?",
      "answer_type": "yes_no"
    }
  ]
}

Hard constraints:
- `sub_questions` must contain exactly <|NUM_SUBQUESTIONS|> items.
```

```
- IDs must be `q1` to `q<|NUM_SUBQUESTIONS|>` in order.
- Every `question` must end with a question mark.
- Every `question` must be answerable by `Yes` or `No`.
- Every `answer_type` must be exactly `yes_no`.
- Returning 1, 2, 3, 4, or more than <|NUM_SUBQUESTIONS|> sub-
    questions is invalid.

Example input prompt:
A red fox leaps across a mossy fallen log in a misty forest while the
    camera tracks smoothly from
left to right.

Example output:
{
  "overall_question": "Does the generated video faithfully depict the
    main scene elements and
  motion described in the prompt?",
  "sub_questions": [
    {"id": "q1", "question": "Is a red fox visible in the video?", "
    answer_type": "yes_no"},
    {"id": "q2", "question": "Is the fox moving in a leaping or
    jumping motion?",
     "answer_type": "yes_no"},
    {"id": "q3", "question": "Is a mossy fallen log visible in the
    scene?",
     "answer_type": "yes_no"},
    {"id": "q4", "question": "Is the environment a misty forest?", "
    answer_type": "yes_no"},
    {"id": "q5", "question": "Does the camera track smoothly from left
     to right?",
     "answer_type": "yes_no"}
  ]
}

Now analyze this prompt:

<|PROMPT|>
```

The decomposition is generated deterministically with a Qwen3.5-9B language model using temperature 0 and top-p of 1.0. We request exactly five sub-questions for the main experiments. If the output is invalid JSON or violates the schema, the script retries with a repair prompt that repeats the original prompt, the invalid output, and the validation error.

B.6 VLM REWARD PROMPT AND REFERENCE ANSWER

After decomposition, the sub-questions are formatted into a structured video QA prompt. This is the prompt text given to the VLM together with sampled video frames:

VLM reward prompt

```
You are a careful video question-answering assistant.
Use only evidence grounded in the video, and do not guess unseen
    details.

Given an AI-generated video, answer the following atomic video-
    grounded questions using only evidence from
the video.

Rules:
- Answer each question independently.
- Use concise canonical phrases.
```

```
- Do not add explanations.
- For "yes_no" questions, answer with "Yes" or "No" only.
- Return valid JSON only.

Output format:
{"sub_answers": [{"id": "q1", "answer": "..."}]}

Sub-questions:
- {"id": "q1", "question": "<question 1>", "answer_type": "yes_no"}
- {"id": "q2", "question": "<question 2>", "answer_type": "yes_no"}
- {"id": "q3", "question": "<question 3>", "answer_type": "yes_no"}
- {"id": "q4", "question": "<question 4>", "answer_type": "yes_no"}
- {"id": "q5", "question": "<question 5>", "answer_type": "yes_no"}
```

The reference answer is appended as the assistant message and is used only for teacher-forced scoring:

Teacher-forced reference answer

```
{"sub_answers": [
  {"id": "q1", "answer": "Yes"},
  {"id": "q2", "answer": "Yes"},
  {"id": "q3", "answer": "Yes"},
  {"id": "q4", "answer": "Yes"},
  {"id": "q5", "answer": "Yes"}
]}
```

The reward for a slot is the mean target-answer log-probability over the answer-value tokens in the reference answer. Slot rewards are averaged into the rollout reward before group-relative normalization. Credit weights come from the absolute gradient of each slot reward with respect to the processed VLM video-frame input, normalized across time and space for spatial routing (across time for frame-only routing). The maps remain question-conditioned but share one rollout-level advantage. All slots are evaluated in one teacher-forced sequence; later slots are conditioned on preceding reference tokens, not statistically independent answers. Answer-token predictions from that same forward pass are used only for monitoring: they neither select the fixed reference target nor enter the main reward. Autoregressive free-generation scoring is a separate ablation.

B.7 Mathematical interpretation of token-credit routing

This section gives conditional intuitions, not guarantees, for the weighted-log-probability update. Equation (5) is not the exact likelihood ratio of the denoising trajectory, and each claim below holds only under the stated assumptions.

Consider one rollout i, one denoising step t, and one video-grounded check j. Ignoring clipping, the weighted-log-probability loss term is

$$\mathcal{L}_{i,t,j}^{\text{unclip}} = -\hat{A}_i \rho_{i,t,j}(\theta), \qquad \rho_{i,t,j}(\theta) = \exp\left(\sum_f w_{i,j,f}\left(u_{i,t,f}^{\theta} - u_{i,t,f}^{\theta_{\text{old}}}\right)\right). \tag{B4}$$

Because the old-policy log-probabilities and the frame weights are detached from the generator update, the gradient is

$$\nabla_\theta \mathcal{L}_{i,t,j}^{\text{unclip}} = -\hat{A}_i \rho_{i,t,j}(\theta) \sum_f w_{i,j,f} \nabla_\theta u_{i,t,f}^{\theta}. \tag{B5}$$

Thus the contribution of frame f to the policy-gradient update routed through check j is proportional to $w_{i,j,f}$. The shared advantage $\hat{A}_i$, computed from the average reward across checks, determines the sign and overall magnitude of the update: positive-advantage samples are reinforced and negative-advantage samples are suppressed, with the largest change assigned to high-credit frames. The map is a detached routing distribution, not an independent question-level advantage.

The first-order policy-improvement view explains when this routing is useful. For a small update from the old policy, the expected reward improvement can be approximated by

$$\Delta J^{(1)} \approx \mathbb{E}\left[A(x)\Delta \log \pi(x)\right]. \tag{B6}$$

If the video log-probability decomposes into frame-level changes $\Delta \log \pi_f$, standard scalar GRPO uses the full or uniformly averaged trajectory change, while TVRL uses the credit-weighted local surrogate

$$\Delta \log \pi_{\text{tvrl}}(x) = \sum_{f=1}^{F} w_f \Delta \log \pi_f. \tag{B7}$$

This surrogate is biased relative to the exact trajectory likelihood, but it is beneficial when the weight vector is aligned with the frames that can actually change the reward.

To make this condition explicit, let $c_f \geq 0$ denote the true local reward sensitivity of frame f, and let u_f denote the update magnitude assigned to that frame. Under a fixed KL-like update budget $\sum_f u_f^2 \leq B$, a frame-level linear model gives

$$\Delta J^{(1)}_{\text{frame}} \approx \sum_{f=1}^{F} c_f u_f \leq \sqrt{B}\,\|c\|_2, \tag{B8}$$

where equality is achieved when $u_f \propto c_f$. By contrast, a uniform allocation gives

$$\Delta J^{(1)}_{\text{uniform}} = \sqrt{\frac{B}{F}} \sum_{f=1}^{F} c_f \leq \sqrt{B}\,\|c\|_2, \tag{B9}$$

with equality only when the reward sensitivity is uniform over frames. This is an idealized allocation problem: it identifies $u_f \propto c_f$ as optimal, but it does not show that unsigned VLM sensitivity induces such updates through a shared-parameter generator, nor that positive alignment alone beats uniform allocation.

For variance, compare against the uniform average $g_{\text{unif}} = \frac{A}{F}\sum_f h_f$ with $h_f = \nabla_\theta \log \pi_\theta(x_f)$, which has the same scale as

$$g_{\text{tvrl}} = A \sum_{f=1}^{F} w_f \nabla_\theta \log \pi_\theta(x_f) \tag{B10}$$

For independent h_f with $\operatorname{tr}\operatorname{Cov}(h_f) = \sigma_f^2$, weighting lowers variance only if $\sum_f w_f^2 \sigma_f^2 < F^{-2}\sum_f \sigma_f^2$. With equal σ_f it cannot, since $\sum_f w_f^2 \geq 1/F$; a reduction requires that low-credit frames carry more reward-irrelevant noise than high-credit ones.

The same variance argument also explains why arbitrarily fine credit need not be optimal. Let the estimated dense credit be $\hat{c}_g = c_g + \varepsilon_g$, where c_g is the reward-relevant local sensitivity and ε_g is model- and sample-dependent estimation noise. A small aggregation window introduces little spatial bias but retains more of ε_g; after nonnegative normalization, isolated responses can produce a concentrated weight vector. For weakly correlated unit-level policy gradients $h_g = \nabla_\theta \log \pi_\theta(x_g)$, the conditional covariance approximately satisfies

$$\operatorname{tr}\operatorname{Cov}\left(A\sum_g w_g h_g \mid w\right) \approx A^2 \sum_g w_g^2 \operatorname{tr}\operatorname{Cov}(h_g). \tag{B11}$$

When the unit-level variances are comparable, a peaky 1×1 map has a smaller effective number of credited units, $N_{\text{eff}}(w) = 1/\sum_g w_g^2$, and can yield a higher-variance update. Increasing the spatial window averages local estimation noise and typically raises N_{eff}, but an overly large window introduces localization bias by mixing distinct semantic regions. The preferred window therefore balances estimator variance against spatial bias; the analysis predicts an intermediate optimum rather than monotonic improvement as the credit map becomes finer.

Finally, PPO clipping does not change the meaning of the weights. Inside the unclipped region, Equation (B5) gives the exact local gradient of the surrogate. When clipping is active, the gradient

from that sample–check pair is bounded or removed to avoid an overly large policy update, but the frame weights still specify where the active policy-gradient signal is routed.

This analysis is conditional: it does not establish that a VLM gradient is a perfect causal explanation of video semantics. The gradient is a local, model-dependent sensitivity measure, and its absolute value captures importance rather than sign. The practical claim is that when this sensitivity aligns with prompt-relevant temporal evidence, token credit concentrates the GRPO update on the frames where changes most likely affect the reward.

B.8 Unsmoothed cell-level credit

The main-text optimization comparison focuses on the consistent w/o-credit $\rightarrow$ frame $\rightarrow 7\times7 \rightarrow 3\times3$ progression. Figure B1 additionally reports the unsmoothed 1×1 control. This extreme does not continue the trend: it underperforms the spatially aggregated variants with Qwen3.5-9B, and its VideoScore2 reward drops sharply, indicating evaluator-dependent optimization instability. A held-out Qwen3.5-9B evaluation is consistent with this observation: 1×1 obtains 55.10 Overall, below frame credit (56.47), 7×7 (57.08), and 3×3 (57.69).

The raw VLM-gradient map is a local sensitivity estimate rather than a noise-free causal attribution. Without spatial aggregation, isolated high-frequency responses can survive normalization and assign disproportionate credit to a few latent cells. This lowers the effective number of credited units and can increase update variance, as described by Equation (B11). PPO clipping limits excessive likelihood-ratio contributions but cannot correct a persistently mislocalized routing direction. The more severe VideoScore2 failure suggests that robustness at cell-level resolution also depends on the structure of the evaluator's reward sensitivity; it implies neither that question-level advantages are required nor that token credit is ineffective.

Together with the held-out result above, Figure B1 shows that removing spatial aggregation does not extend the improvement obtained by moving from frame-level to spatially aggregated credit. The complementary 3×3 versus 7×7 residual diagnostic appears in Figure 8 of the main paper, and Figure B2 shows two further examples, with the strongest residuals near the pen barrel and writing region and near the white headband.

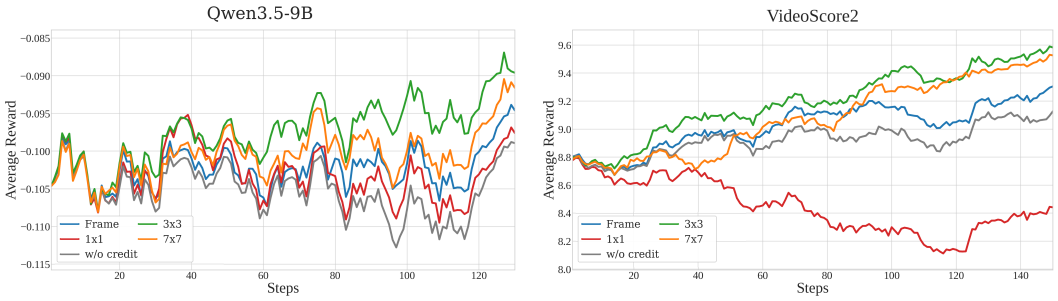

Figure B1: Supplementary optimization traces including the unsmoothed 1×1 credit map for Qwen3.5-9B (left) and VideoScore2 (right). The cell-level variant does not continue the improvement obtained by progressively localizing spatially aggregated credit and is particularly unstable with VideoScore2.



Is a hand holding a black fountain pen with a gold-rimmed barrel visible writing on a sheet of paper?



Does the woman have long, raven-black hair held back by a simple white washi headband?

Figure B2: **Additional spatial residual diagnostics.** The positive $3 \times 3 - 7 \times 7$ residual for two further question-conditioned maps, shown over five sampled frames (queried evidence in red). At high-credit frames, the strongest residuals lie near the pen and gold-rimmed barrel and near the white headband. These per-example normalized sensitivity maps are not causal masks.