GRAFTOOD: GRAFTING OUT-OF-DISTRIBUTION EXEMPLARS INTO ADVANTAGE-WEIGHTED RECTIFICATION FOR VIDEO REWARD OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video reward optimization can increase proxy rewards while reducing video dynamics, a mismatch we identify as *motion collapse*, including under a video-level motion-quality reward. We propose **GraftOOD**, which replaces part of the online rollouts in selected prompt groups with out-of-distribution video exemplars. A shared velocity-regression objective combines group-relative reward rectification with sequence-level exemplar supervision, requiring neither exemplar generation trajectories nor an additional supervised-loss coefficient. We characterize its pointwise optimal field as a source-posterior-weighted combination of the two source-specific target fields. We construct **GraftSet-Video**, containing 9,588 format-matched exemplars for 2,397 prompts, to supply external supervision. On Wan2.1-1.3B, GraftOOD raises Dynamic Degree from 4.86 to 44.41 relative to reward-only FlowAWR, while HPSv3 decreases from 8.10 to 7.32 and VisionReward remains close. At the main setting, GraftOOD leads nine of ten metrics among three exemplar-augmented methods. Ablations of exemplar frequency and within-group count reveal dynamics–reward trade-offs rather than uniform gains from increasing exemplar usage. These results support external exemplar supervision as a complement to scalar reward feedback, without equating higher measured dynamics with coherent, visually plausible motion in the generated videos.

1 INTRODUCTION

Reinforcement learning has become a central component of generative-model post-training, enabling learning from human preferences (Christiano et al., 2017; Ouyang et al., 2022) and verifiable outcomes (Shao et al., 2024). Unlike supervised imitation alone, it lets a model sample its own outputs, receive feedback, and concentrate updates on behaviors that satisfy the specified criteria. This principle extends to visual generation, where learned evaluators guide improvements in aesthetic quality and prompt correspondence (Black et al., 2024; Fan et al., 2023; Xue et al., 2025; Zheng et al., 2025). These advances depend not only on the optimization algorithm, but also on how well the reward represents the preferences that define the desired outputs.

For video generation, a proxy reward must account for both individual-frame quality and how content evolves over time. Optimizing this proxy favors high-scoring outputs, which need not satisfy all of these preferences. In our experiments, optimizing frame-aggregated HPSv3 (Ma et al., 2025; Team et al., 2025b) raises its score while reducing Dynamic Degree (Huang et al., 2024). Optimizing VideoAlign's Motion Quality (Liu et al., 2026b) likewise raises its reward while reducing Dynamic Degree. We refer to this mismatch between rising rewards and declining video dynamics as *motion collapse*. These observations are consistent with reward exploitation: a policy may increasingly favor high-reward outputs whose temporal behavior is undesirable.

Existing methods improve reward optimization through policy gradients or regression objectives (Liu et al., 2026a; Li et al., 2025; He et al., 2026), but improving a proxy objective does not by itself resolve its mismatch with the desired outputs. Combining rewards is one response, yet it requires choosing which attributes to score and how to balance them. In our experiments, adding a RAFT-based motion reward (Teed & Deng, 2020) to HPSv3 mitigates but does not prevent the decline in Dynamic Degree. Rather than relying solely on further reward design, external videos can provide

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: **Reward-only optimization versus GraftOOD.** Same-prompt sequences from reward-only optimization (left) and GraftOOD (right), with frames in temporal order.

examples of outputs we want the policy to learn from. This raises the question: **Can an external exemplar distribution guide online reward optimization toward desired video attributes that the reward does not adequately constrain?**

DiffusionNFT (Zheng et al., 2025) optimizes rewards through velocity regression rather than likelihood ratios over generation trajectories. FlowAWR (Fu et al., 2026) further derives advantage-weighted velocity rectification. This regression view provides a shared interface for online reward optimization and external supervision: both update the same velocity network, but they differ in the sample sources and regression targets that define each update.

We introduce **GraftOOD**, which replaces part of the online rollouts in selected prompt groups with out-of-distribution (OOD) video exemplars. Online targets retain group-relative reward rectification, while exemplars provide sequence-level supervision without their generating trajectories. Both enter one regression objective without an additional supervised-loss coefficient. We characterize its pointwise optimal field as a source-posterior-weighted combination of the online target field and the exemplar reference field (Figure 2, top). This analysis describes the regression target, rather than guaranteeing properties of generated videos. Figure 1 gives same-prompt qualitative comparisons of video sequences generated with GraftOOD and reward-only optimization over multiple frames.

We construct **GraftSet-Video** from public video reward-model resources, supplementing selected prompt–video pairs with outputs from external generators, including Seedance and Kling. The corpus contains 9,588 videos for 2,397 prompts outside the original policy-training set, with four exemplars per prompt and the policy's resolution and frame count.

On Wan2.1-1.3B, GraftOOD uses an exemplar-prompt ratio of 0.167 and four exemplars per group in the main comparison. Relative to FlowAWR, it raises Dynamic Degree from 4.86 to 44.41, while HPSv3 decreases from 8.10 to 7.32 and VisionReward remains close. GraftOOD leads nine of ten metrics in this comparison with FlashOOD and NFTOOD. Ablations of both exemplar frequency and within-group count reveal dynamics–reward trade-offs, not uniform improvements from increasing exemplar usage. Higher Dynamic Degree alone does not establish better motion quality.

Our contributions are as follows:

- We identify *motion collapse* in video reward optimization: increasing proxy rewards can coincide with declining video dynamics, even with a video-level motion-quality reward. These observations motivate sequence-level supervision beyond scalar reward feedback.
- We propose GraftOOD, which grafts OOD video exemplars into group-relative online optimization through a shared velocity-regression objective, without an additional supervised-loss coefficient. We characterize its pointwise optimal field as a source-posterior-weighted combination of the two source-specific target fields.

- We construct GraftSet-Video, comprising 9,588 exemplars for 2,397 prompts, and evaluate exemplar-augmented optimization on Wan2.1-1.3B. GraftOOD improves dynamics over reward-only FlowAWR and leads nine of ten metrics among the three OOD methods at the main setting, while ratio and count ablations reveal dynamics–reward trade-offs.

2 BACKGROUND

2.1 TRAJECTORY-BASED POLICY OPTIMIZATION

Flow models generate samples by integrating the deterministic ODE $\mathrm{d}x_t = v_\theta(x_t, c, t)\,\mathrm{d}t$. For policy-gradient training, FlowGRPO (Liu et al., 2026a) instead adopts a marginal-preserving SDE, whose discretization provides tractable transition likelihoods.

For each prompt c, the terminal rewards of G on-policy trajectories are normalized using the prompt-group mean $\overline{R}_c$ and the reward standard deviation over the current rollout batch to form advantages:

$$\mathcal{A}^{(i)} = \frac{R(c, x_1^{(i)}) - \overline{R}_c}{\mathrm{Std}(R) + \epsilon}. \tag{1}$$

These advantages weight the trajectory log-probabilities in the on-policy surrogate objective:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\mathbb{E}_{c,\{\tau^{(i)}\}\sim p_\theta}\left[\frac{1}{G}\sum_{i=1}^{G}\mathcal{A}^{(i)}\sum_{k=0}^{K-1}\log p_\theta(x_{t_{k+1}}^{(i)} \mid x_{t_k}^{(i)}, c)\right]. \tag{2}$$

where the likelihood terms are evaluated along the sampled SDE paths.

2.2 REWARD-GUIDED VELOCITY REGRESSION

Regression-based methods instead incorporate rewards into velocity regression without evaluating trajectory likelihoods. Under Rectified Flow (Lipman et al., 2022; Liu et al., 2023), a terminal video latent x_1 and fresh noise $x_0 \sim p_0 = \mathcal{N}(0, I)$ define the training state and conditional velocity target:

$$x_t = (1-t)x_0 + tx_1, \qquad u_t(x_t \mid x_1) = x_1 - x_0 = \frac{x_1 - x_t}{1-t}. \tag{3}$$

DiffusionNFT (Zheng et al., 2025) uses the normalized terminal reward $r \in [0,1]$ to weight regression toward positive and negative corrections of a fixed reference field v^{old}. For fixed $\beta > 0$, its objective can be rewritten, omitting spatiotemporal arguments, as

$$\mathcal{L}_{\mathrm{NFT}}(\theta) = \beta^2\mathbb{E}\Big[r\left\|v_\theta - \left(v^{\mathrm{old}} + \frac{u_t - v^{\mathrm{old}}}{\beta}\right)\right\|_2^2 + (1-r)\left\|v_\theta - \left(v^{\mathrm{old}} - \frac{u_t - v^{\mathrm{old}}}{\beta}\right)\right\|_2^2\Big]. \tag{4}$$

Starting from the optimal terminal distribution under KL-regularized reward maximization, FlowAWR (Fu et al., 2026) derives its intermediate marginals and corresponding velocity field:

$$v^\star(x_t, c, t) = v^{\mathrm{old}}(x_t, c, t) + \mathbb{E}_{x_1\sim p^{\mathrm{old}}(\cdot|x_t,c)}\left[\frac{\exp(R(c,x_1)/\gamma)}{\phi(x_t,c,t)}(u_t - v^{\mathrm{old}})\right]. \tag{5}$$

Since $\mathbb{E}_{p^{\mathrm{old}}(\cdot|x_t,c)}[u_t - v^{\mathrm{old}}] = 0$, subtracting the normalized weight's mean yields the advantage

$$\mathcal{A}_t(x_1, x_t, c) = \frac{\exp(R(c,x_1)/\gamma)}{\phi(x_t,c,t)} - 1, \tag{6}$$

with $\gamma > 0$ and $\phi(x_t, c, t) = \mathbb{E}_{x_1'\sim p^{\mathrm{old}}(\cdot|x_t,c)}[\exp(R(c,x_1')/\gamma)]$. Following the reduction in conditional flow matching (Lipman et al., 2022), regression toward $v^\star$ uses sample targets:

$$\mathcal{L}_{\mathrm{AWR}}(\theta) = \mathbb{E}_{\substack{x_1\sim p^{\mathrm{old}}(\cdot|c)\\ t,\,x_0\sim p_0}}\left[\left\|v_\theta - \left(v^{\mathrm{old}} + \mathcal{A}_t(x_1,x_t,c)\left(u_t - v^{\mathrm{old}}\right)\right)\right\|_2^2\right], \tag{7}$$

where GRPO-style estimation gives $\mathcal{A}_t \approx \widehat{\mathcal{A}}_0^{(i)} = \exp(R_i/\gamma)/[G^{-1}\sum_{j=1}^{G}\exp(R_j/\gamma)] - 1$.

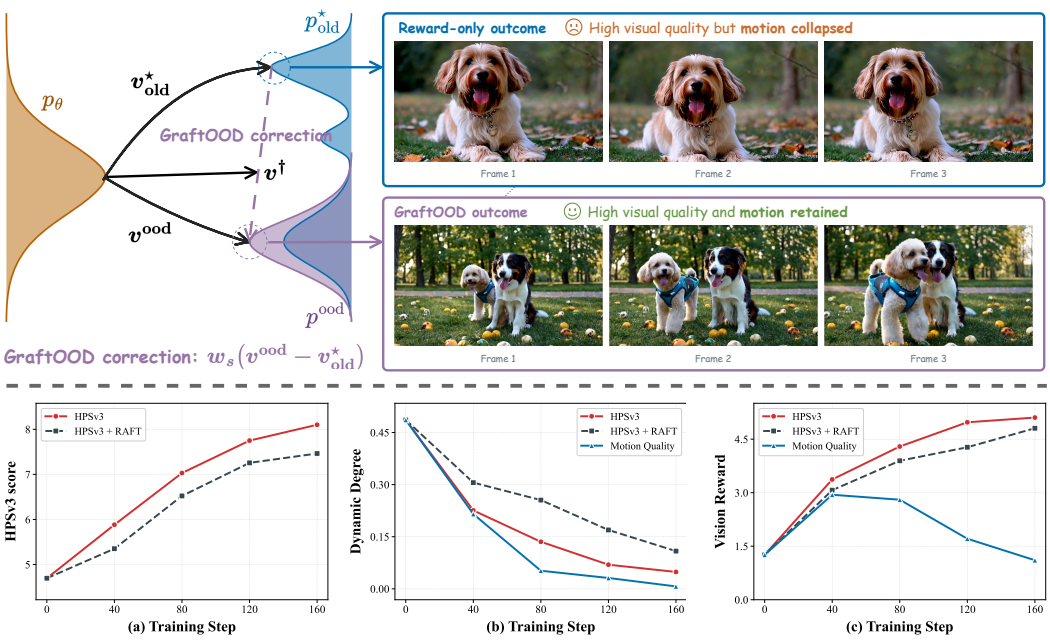


Figure 2: **GraftOOD motivation and overview. Top:** exemplar-guided correction $w_s(v^{\text{ood}} - v^\star_{\text{old}})$ and motion examples. **Bottom: (a)** HPSv3, **(b)** Dynamic Degree, and **(c)** VisionReward during reward optimization. Adding RAFT to HPSv3 mitigates but does not prevent declining dynamics.

3 Video Reinforcement with OOD Distribution Grafting

3.1 Problem Setup

Consider a video policy conditioned on a prompt c and updated using a terminal reward $R(c, x_1)$. Online samples provide training experience, while the reward determines which outputs to encourage or suppress. As a proxy for the desired video attributes, the reward may also favor undesirable outputs. Optimizing it can then concentrate the policy on a high-reward region that does not reflect our preferences for visual appearance and temporal behavior.

Figure 2(a,b) illustrates this mismatch: optimizing HPSv3 raises its score while Dynamic Degree declines. Optimizing VideoAlign Motion Quality likewise raises its reward while reducing Dynamic Degree (Appendix B.3). Adding a RAFT-based motion reward mitigates but does not prevent the decline observed with HPSv3 in the same panel. Meanwhile, Figure 2(c) shows that HPSv3 training increases VisionReward (Xu et al., 2026), an evaluator not used for training. These results motivate **supplementing reward optimization rather than replacing it**: improvement on another evaluator coexists with loss of dynamics in the videos produced during reward optimization.

Additional reward terms revise how online samples are scored, requiring choices about the attributes to measure and their relative weights. We instead introduce external videos as a distributional reference alongside reward feedback. They supply sequences for the policy to learn from without first generating those examples itself. The selected exemplars provide direct sequence-level targets to guide policy learning toward desired video attributes.

We construct the OOD exemplar distribution $p^{\text{ood}}(x_1 \mid c)$ from prompt–video pairs selected from public video reward-model resources (Liu et al., 2026b; Xu et al., 2026; He et al., 2024; 2025; Tong et al., 2025), supplemented with outputs from external generators such as Seedance and Kling. These videos are selected to provide higher-quality references for the base policy. Their paired prompts lie outside the original policy-training set, and the videos match the policy's training format. OOD is defined relative to this policy setting, not necessarily to the reward model's training data. The policy also generates fresh rollouts on these prompts for reward-relative updates. The exemplars provide clean terminal latents, but not their generating trajectories. *How can we incorporate these external video references into online reward optimization, so that policy updates are guided by both reward feedback and the sequence-level attributes of the selected exemplars?*

3.2 GROUP-WISE EXEMPLAR GRAFTING

Trajectory-based optimization with exemplar supervision. These terminal-only exemplars cannot be inserted directly as on-policy samples in Equation 2. We therefore supplement online policy optimization with supervised fine-tuning (SFT), which constructs regression inputs by adding noise to the encoded exemplar latents. We use Flash-GRPO's single-timestep scheme (He et al., 2026) to train the online and exemplar objectives at a matched noise level. For each prompt group s, we randomly select a timestep t_s, compute the online policy loss from the corresponding stochastic transition, and construct the exemplar SFT loss at the same noise level. Specifically, each encoded exemplar $x_1^{(j)}$ is independently noised as $x_{t_s}^{(j)} = (1-t_s)x_0^{(j)} + t_s x_1^{(j)}$, with $x_0^{(j)} \sim p_0$. The network takes $x_{t_s}^{(j)}$ as input and regresses $u_{t_s}^{(j)} = u_{t_s}(x_{t_s}^{(j)} \mid x_1^{(j)})$, defined in Equation 3. Using the on-policy form of Equation 2, we combine the single-transition policy objective with exemplar SFT at the shared timestep t_s:

$$\widehat{\mathcal{L}}_s = -\frac{1}{m_s}\sum_{i=1}^{m_s} \mathcal{A}^{(i)} \log p_\theta\Big(x_{t_s+1}^{(i)} \mid x_{t_s}^{(i)}, c_s\Big) + \frac{\lambda_{\text{ood}}}{n_{\text{ood}}}\sum_{j=1}^{n_{\text{ood}}} \Big\| v_\theta(x_{t_s}^{(j)}, c_s, t_s) - u_{t_s}^{(j)} \Big\|_2^2. \tag{8}$$

For the m_s online rollouts, $x_{t_s+1}^{(i)}$ denotes the latent obtained from $x_{t_s}^{(i)}$ by one discretized SDE denoising step. The advantages $\mathcal{A}^{(i)}$ are computed exclusively from online-rollout rewards using the normalization in Equation 1. The coefficient λ_{ood} controls the relative weight of exemplar SFT. We refer to this exemplar-augmented formulation as **FlashOOD**.

Reinforcement learning through terminal fine-tuning. Off-policy RL through terminal fine-tuning provides a direct regression interface for these exemplars. In the schemes considered here (Zheng et al., 2025; Fu et al., 2026), sampling uses an old policy whose parameters are an exponential moving average (EMA) of the trainable parameters. Training re-noises the collected clean latents for reward-guided velocity regression. For an external supervision term, terminal sampling in the outer expectation changes from $x_1 \sim p^{\text{old}}(\cdot \mid c)$ to $p^{\text{ood}}(\cdot \mid c)$, while the online term is retained. When using normalized rewards, we assign the OOD references the theoretically maximal reward.

For example, DiffusionNFT uses the reward-weighted positive and negative corrections in Equation 4. Let $\widehat{\mathcal{L}}_{\text{NFT},s}$ denote its empirical average over the m_s old-policy rollouts. For exemplars, $r = 1$ removes the negative term. With $v^{\text{old},(j)} = v^{\text{old}}(x_{t_s}^{(j)}, c_s, t_s)$, we combine the online and exemplar mean losses with equal weights at the shared t_s:

$$\widehat{\mathcal{L}}_{\text{NFT},s}^{\text{ood}} = \widehat{\mathcal{L}}_{\text{NFT},s} + \frac{1}{n_{\text{ood}}}\sum_{j=1}^{n_{\text{ood}}} \Big\| (1-\beta)v^{\text{old},(j)} + \beta v_\theta(x_{t_s}^{(j)}, c_s, t_s) - u_{t_s}^{(j)} \Big\|_2^2. \tag{9}$$

The online objective weights both corrections by r_i and $1 - r_i$, whereas exemplars receive only the positive correction. And under the default setting $\beta = 1$ (Zheng et al., 2025), the OOD term reduces to standard SFT, regressing $v_\theta(x_{t_s}^{(j)}, c_s, t_s)$ directly toward $u_{t_s}^{(j)}$. We refer to this DiffusionNFT formulation with positive-only exemplar supervision as **NFTOOD**.

Group-relative velocity rectification and exemplar grafting. Each online sample provides its own residual direction $u_t^{(i)} - v^{\text{old},(i)}$, whose contribution to the regression target is controlled by its group-relative coefficient $\widehat{\mathcal{A}}_0^{(i)}$. Assume an OOD reference policy consistent with p^{ood} and the same forward noising path, with velocity field $v^{\text{ood}}(x_t, c, t)$. Applying Equation 5 to this reference without centering the normalized reward coefficient gives

$$v_{\text{ood}}^\star = v^{\text{ood}} + \mathbb{E}_{p^{\text{ood}}(\cdot|x_t,c)}\left[\mathcal{A}_t^{\text{ood}}(u_t - v^{\text{ood}})\right]. \tag{10}$$

We regard the selected OOD distribution as a desired reference, accepting its exemplars without further training-time reward differentiation. Assigning a common reference reward $R^\star$ gives

$$\mathcal{A}_t^{\text{ood}}(x_1, x_t, c) := \frac{\exp(R^\star/\gamma)}{\mathbb{E}_{x_1' \sim p^{\text{ood}}(\cdot|x_t,c)}[\exp(R^\star/\gamma)]} = 1. \tag{11}$$

For an exemplar prompt c_s, let $\delta = n_{\text{ood}}/G$ be the exemplar fraction in a group of size G. Following the conditional flow matching reduction used in Equation 7, we obtain our **GraftOOD** objective by

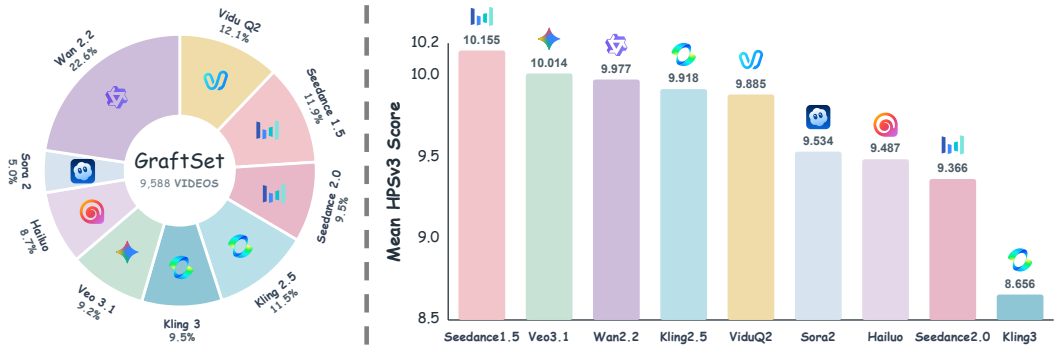


Figure 3: **GraftSet-Video composition and scores.** Source proportions among 9,588 exemplars (left) and mean HPSv3 by source, ranked in descending order (right). Colors match across panels.

combining sample-level regression under the two terminal distributions:

$$\begin{aligned}\mathcal{L}_{\text{GraftOOD}}(\theta \mid c_s) = (1-\delta)\mathbb{E}^{c_s}_{\text{old}}\left[\left\|v_\theta - \left[v^{\text{old}} + \widehat{\mathcal{A}}_0(u_t - v^{\text{old}})\right]\right\|_2^2\right] \\ + \delta\mathbb{E}^{c_s}_{\text{ood}}\left[\left\|v_\theta - \left[v^{\text{ood}} + \mathcal{A}^{\text{ood}}_t(u_t - v^{\text{ood}})\right]\right\|_2^2\right].\end{aligned} \tag{12}$$

where the two expectations sample terminals from $p^{\text{old}}(\cdot \mid c_s)$ and $p^{\text{ood}}(\cdot \mid c_s)$ under the same re-noising timestep. In the OOD term, $\mathcal{A}^{\text{ood}}_t = 1$ cancels the reference field, making the exemplar loss directly computable from re-noised exemplars. And the online term retains group-relative rectification through $\widehat{\mathcal{A}}_0$, which is estimated from online rollouts only. Then **what velocity field do online reward rectification and exemplar supervision jointly define on an exemplar prompt?**

Proposition 1 (Exemplar-guided velocity rectification)**.** *The pointwise optimal velocity field of* $\mathcal{L}_{\text{GraftOOD}}(\theta \mid c_s)$ *satisfies*

$$v^\dagger = (1-w_s)v^\star_{\text{old}} + w_s v^{\text{ood}}, \tag{13}$$

where $v^\star_{\text{old}}$ *is the reward-rectified old-policy target mean conditional on* (x_t, c_s, t)*, including group randomness. The proof is given in Appendix A.2. The exemplar-source posterior is*

$$w_s(x_t, t) = \frac{\delta p^{\text{ood}}_t(x_t \mid c_s)}{(1-\delta)p^{\text{old}}_t(x_t \mid c_s) + \delta p^{\text{ood}}_t(x_t \mid c_s)}. \tag{14}$$

Relative to the online target under the same group-sampling scheme, exemplars contribute $w_s(v^{\text{ood}} - v^\star_{\text{old}})$. The target therefore combines the reward-rectified old-policy field with an external reference rather than relying on reward rectification alone. Their relative weight depends on both the exemplar fraction and the noised-state densities.

3.3 Practical Implementation

GraftSet-Video. We construct **GraftSet-Video** with exactly four video exemplars per prompt, yielding 9,588 videos for 2,397 prompts. We first select videos from public video reward-model training and evaluation resources, including those accompanying Liu et al. (2026b); Xu et al. (2026); He et al. (2024; 2025); Tong et al. (2025), using HPSv3 scores (Ma et al., 2025). Keeping the selected videos, we supplement prompts with fewer than four exemplars using additional outputs from external generators, including Seedance (Seedance et al., 2025; 2026) and Kling (Team et al., 2025a), to meet this four-exemplar requirement. These generators are chosen to provide higher-quality video references for the base policy beyond its own online rollouts.

HPSv3-based selection is performed offline, inherits HPSv3's limited motion sensitivity, and does not enter training-time advantage estimation. Videos contain 81 frames at 832×480 and 16 fps, encoded with the policy's VAE without further resampling. The prompts are disjoint from the original 19,700 training prompts and 300 evaluation prompts. Source resources may overlap reward-model training data; OOD refers to the policy setting above. Figure 3 shows the source composition and the mean HPSv3 score for each source; Appendix B.1 provides dataset details.

Table 1: HPSv3 optimization of Wan2.1-1.3B with CFG-enabled evaluation. OOD methods use $\rho = 0.167$ and $n_{\text{ood}} = 4$. All metrics are higher-is-better (definitions in Appendix B.2). Bold marks column bests among FlashOOD, NFTOOD, and GraftOOD only.

Method	AQ	IQ	SC	OC	DD	VR	VQ	MQ	TA	HPSv3
Wan2.1 T2V 1.3B	65.63	66.79	94.58	88.70	48.61	1.26	-0.59	-0.34	-0.66	4.70
Flash-GRPO	67.46	68.10	95.85	90.36	28.67	1.65	-0.69	-0.43	-0.59	5.11
DiffusionNFT	68.30	71.50	97.49	90.9	14.24	5.20	-0.09	0.21	-0.51	7.23
FlowAWR	69.18	71.01	98.48	92.53	4.86	5.11	0.06	0.47	-0.65	8.10
FlashOOD	65.87	67.91	94.68	88.29	**58.26**	1.65	-0.62	-0.47	-0.50	5.31
NFTOOD	66.10	70.52	96.99	87.91	41.12	3.90	-0.45	-0.25	-0.47	6.55
GraftOOD	**67.21**	**71.26**	**98.20**	**90.10**	44.41	**5.08**	**-0.25**	**-0.07**	**-0.45**	**7.32**

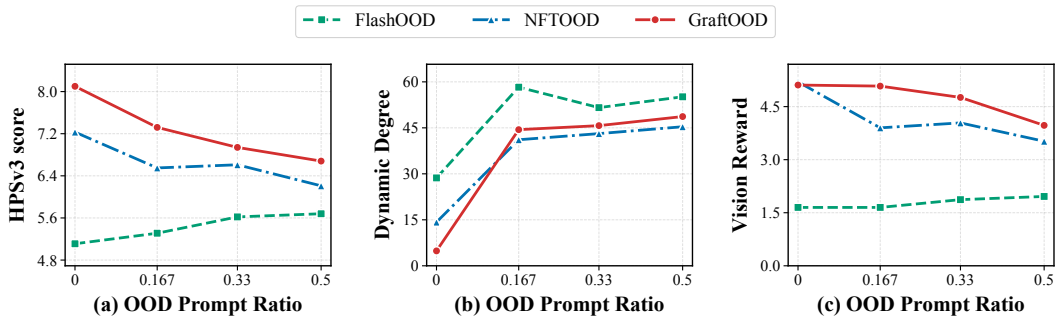


Figure 4: **OOD prompt-ratio sweep.** **(a)** HPSv3, **(b)** DD, and **(c)** VisionReward with $n_{\text{ood}} = 4$ at $\rho > 0$. At $\rho = 0$, FlashOOD, NFTOOD, and GraftOOD use Flash-GRPO, DiffusionNFT, and FlowAWR, respectively, without introducing external exemplars.

Exemplar scheduling and allocation. We control exemplar usage through the OOD prompt ratio ρ and the within-group exemplar count n_{ood}. The ratio ρ determines the fraction of training prompt groups drawn from GraftSet-Video. Each such group combines n_{ood} of the prompt's four stored exemplars with $G - n_{\text{ood}}$ fresh online rollouts. The remaining groups use original training prompts with G online rollouts each and no external exemplar supervision.

4 Experiments

4.1 Experimental Setup

Datasets, Models, and Baselines. Following DanceGRPO (Xue et al., 2025), we use 19,700 training prompts and add exemplars from GraftSet-Video. On Wan2.1-1.3B (Wan et al., 2025), we compare the base model, Flash-GRPO (He et al., 2026), DiffusionNFT (Zheng et al., 2025), and FlowAWR (Fu et al., 2026), alongside their exemplar-augmented variants FlashOOD, NFTOOD, and GraftOOD, as defined by the three training objectives in Section 3.2.

Implementation Details. Training uses 20-step rollouts, classifier-free guidance (CFG) scale 4.5, and group size $G = 8$. NFT/AWR variants use EMA sampling and CFG-free updates. We train rank-32 attention LoRA adapters with AdamW at learning rate 10^{-4}, freezing the backbone. The main OOD comparison uses $\rho = 0.167$ and $n_{\text{ood}} = 4$. All methods use 50 sampling steps and CFG scale 4.5 for evaluation, generating 81-frame videos at 832×480 and 16 fps. Further settings, including LoRA scaling, numerical precision, and regularization, appear in Appendix B.2.

Rewards and Evaluation. We optimize HPSv3 (Ma et al., 2025), averaging the top 30% of per-frame scores following LongCat-Video (Team et al., 2025b). We report VBench's Aesthetic Quality, Imaging Quality, Subject Consistency, and Object Class on its official prompt set (Huang et al., 2024). On 300 VidProM prompts, we measure Dynamic Degree, VisionReward (Xu et al., 2026), VideoAlign's Video Quality, Motion Quality, and Text Alignment (Liu et al., 2026b), and HPSv3. Appendix B.2 details the prompt sets, score aggregation, and evaluation protocol.



Figure 5: **Qualitative comparison.** Each prompt block shows Vanilla, FlashOOD, NFTOOD, and GraftOOD from top to bottom, with three frames per method in temporal order.

4.2 Main Results

Relative to FlowAWR, GraftOOD raises Dynamic Degree (DD) from 4.86 to 44.41, while HPSv3 changes from 8.10 to 7.32 and VisionReward from 5.11 to 5.08 (Table 1). Although DD remains below the base model's 48.61, HPSv3 and VisionReward exceed its 4.70 and 1.26. GraftOOD thus retains more dynamics than reward-only FlowAWR while preserving reward and visual-score gains over the base model, despite the remaining gap in measured dynamics. All three exemplar-augmented methods raise DD over their respective online baselines, but HPSv3 rises only for FlashOOD and falls for NFTOOD and GraftOOD under the shared main-comparison configuration.

GraftOOD leads nine of ten metrics among OOD methods, including all four VBench metrics. FlashOOD attains the highest DD (58.26), but the bottom-right basketball example in Figure 5 shows character deformation. DD uses RAFT-estimated optical-flow magnitudes (Huang et al., 2024; Teed & Deng, 2020) without distinguishing coherent motion from deformation. Deformation can therefore inflate DD, offering a possible explanation for FlashOOD's score rather than evidence of better motion. GraftOOD exceeds both OOD baselines on SC, VQ, and MQ at this setting.

4.3 Ablation Studies

OOD prompt ratio at a fixed exemplar count. At $n_{\text{ood}} = 4$, increasing ρ from 0.167 to 0.5 raises GraftOOD's DD from 44.41 to 48.68 but lowers HPSv3 and VisionReward (Figure 4). GraftOOD still leads both reward metrics among OOD methods at every tested positive ratio. The sweep therefore shows a dynamics–reward trade-off, not a uniform improvement.

Exemplar count at a fixed OOD prompt ratio. Among $n_{\text{ood}} \in \{1, 2, 4\}$, one exemplar yields the highest HPSv3 and VisionReward at every tested positive ρ, while four yield the highest DD (Figure 6). At $\rho = 0.167$, reducing the count from four to two lowers DD from 44.41 to 31.41, but raises HPSv3 from 7.32 to 7.60. Two exemplars score below both alternatives on all three metrics at $\rho = 0.5$. Since $G = 8$ is fixed, exemplar count also changes the online sample count for reward

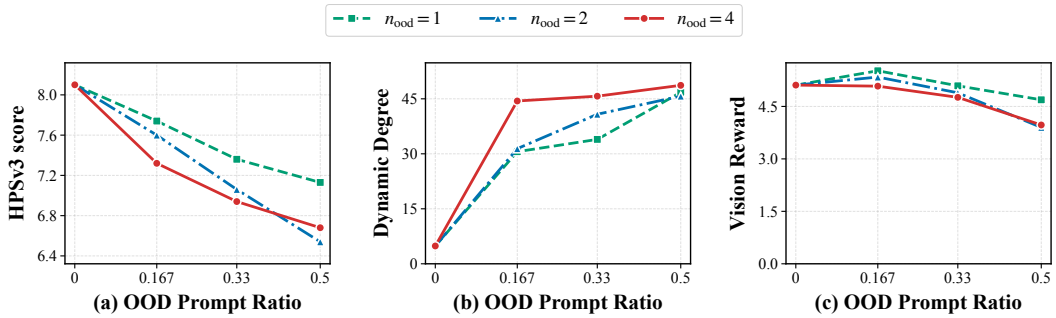


Figure 6: **GraftOOD exemplar-count ablation.** **(a)** HPSv3, **(b)** DD, and **(c)** VisionReward versus ρ for $n_{\text{ood}} \in \{1, 2, 4\}$ ($G = 8$). Curves share one FlowAWR baseline at $\rho = 0$.

normalization: this tests group composition, not exemplar supervision alone. Appendix B.3 reports the full ratio and exemplar-count sweeps with all evaluated quality and reward metrics.

5 Related Work

Reward Alignment for Video Generation. Preference alignment for video generation includes pairwise objectives, such as VideoDPO (Liu et al., 2024), and online reward optimization. DDPO and DPOK formulate diffusion fine-tuning through policy gradients (Black et al., 2024; Fan et al., 2023). Flow-GRPO and DanceGRPO extend group-relative optimization to flow-based visual generation using stochastic sampling (Liu et al., 2026a; Xue et al., 2025), while MixGRPO and Flash-GRPO adopt mixed ODE–SDE sampling and single-timestep updates, respectively (Li et al., 2025; He et al., 2026). Alternatively, DiffusionNFT optimizes reward-weighted regression on the forward process (Zheng et al., 2025), and FlowAWR derives advantage-weighted velocity rectification (Fu et al., 2026). GraftOOD builds on this regression formulation but addresses a different question: how external video exemplars can supplement reward optimization when the proxy inadequately constrains temporal behavior of the generated video sequences.

External Data for Reinforcement Learning. External experience has been incorporated into online RL through different learning signals. RLPD mixes offline transitions with online replay for actor–critic learning (Ball et al., 2023), while LUFFY incorporates expert reasoning trajectories into mixed-policy group-relative optimization (Yan et al., 2026). GraftOOD addresses a different supervision problem: external videos specify desired sequence-level behavior but provide no generating denoising trajectories. This distinction concerns the form of external supervision, not merely its source: each video provides a regression target without a recorded sequence of policy actions. We use these terminal exemplars to construct supervised velocity targets and combine them with reward-relative online targets within prompt groups, excluding exemplars from online reward normalization. Beyond combining data sources, our analysis characterizes the resulting exemplar-guided velocity field. This formulation supplements attributes inadequately constrained by the proxy reward without an additional motion reward or training-time reward-model scoring of external exemplars.

6 Conclusion

We identified motion collapse in video reward optimization, where increasing proxy rewards can coincide with declining video dynamics. We proposed GraftOOD to incorporate external sequence-level supervision into group-relative velocity regression without an additional supervised-loss coefficient. Our analysis characterizes its pointwise optimum as a source-posterior-weighted combination of online and exemplar target fields. With GraftSet-Video, experiments on Wan2.1-1.3B show higher measured dynamics than reward-only FlowAWR, while exemplar-frequency and count ablations reveal trade-offs with HPSv3 and VisionReward. These findings support external exemplars as a complement to scalar reward feedback, rather than a replacement for reward optimization. Exemplar selection and motion-quality evaluation remain limitations: higher Dynamic Degree alone does not establish coherent motion or improved temporal quality in generated videos.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and revision, including wording, LaTeX formatting. AI-assisted checks do not replace author verification. The authors take responsibility for the manuscript, including its technical claims, derivations, experimental results, and references.

ETHICS STATEMENT

GraftOOD uses pretrained video models, learned reward functions, and generated video exemplars. It may inherit biases from these sources, and higher reward scores do not by themselves ensure safe or unbiased outputs. We have not identified additional ethical concerns specific to this study beyond these general risks.

REPRODUCIBILITY STATEMENT

We describe the sample-averaged training objective and sampling assumptions in Appendix A, with the proof of Proposition 1 in Appendix A.2. GraftSet-Video construction and source statistics are described in Section 3.3 and Appendix B.1. We will publicly release GraftSet-Video upon acceptance of the paper to support reproducibility and further research on exemplar-guided video reward optimization. Training configurations, reward scaling, exemplar filtering, and evaluation protocols are provided in Appendix B.2. The motion-quality reward study, additional method comparisons, and exemplar-ratio and count ablations are reported in Appendix B.3. Additional qualitative examples and their prompts appear in Appendix D.

REFERENCES

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 1577–1594. PMLR, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Zheming Fu, Ruizhe He, Wei Shang, Xiaoxiao Ma, Lei Wang, Chang Liu, and Siming Fu. Flowawr: Online adaptive flow reinforcement via advantage-weighted rectification. *arXiv preprint arXiv:2606.30376*, 2026.

Xiaoxuan He, Siming Fu, Zeyue Xue, Weijie Wang, Ruizhe He, Yuming Li, Dacheng Yin, Shuai Dong, Haoyang Huang, Hongfa Wang, et al. Flash-grpo: Efficient alignment for video diffusion via one-step policy optimization. *arXiv preprint arXiv:2605.15980*, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. 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, 2024.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

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

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The eleventh international conference on learning representations*, 2022.

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. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. volume 38, pp. 82155–82192, 2026b.

Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International conference on learning representations (ICLR)*, 2023.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

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

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

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

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025b.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419, 2020.

Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, Shi Qiu, Siwei Han, Kexin Geng, Zhongkai Xue, Yiyang Zhou, Peng Xia, et al. Mj-video: Fine-grained benchmarking and rewarding video preferences in video generation. 2025.

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

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. 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.

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

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *Advances in Neural Information Processing Systems*, 38:117157–117186, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

APPENDIX CONTENTS

A Analysis of Dual-Source Regression . . . 13
A.1 Sample-Averaged Implementation . . . 13
A.2 Sampling Assumptions and Proof . . . 13
B Experimental Details and Additional Results . . . 14
B.1 GraftSet-Video Details . . . 14
B.2 Experimental Setup . . . 14
B.3 Additional Quantitative Results . . . 15
C Limitations and Future Work . . . 17
D More Qualitative Evaluation . . . 18
D.1 Additional Comparisons . . . 18
D.2 Full Prompts for the Qualitative Comparisons . . . 18

A ANALYSIS OF DUAL-SOURCE REGRESSION

A.1 SAMPLE-AVERAGED IMPLEMENTATION

An original training prompt uses G rollouts; an exemplar prompt uses n_{ood} stored videos and $M = G - n_{\mathrm{ood}} \geq 2$ fresh rollouts. All members share the prompt and timestep t_s, with independent fresh noise. Let $\mathcal{I}_s^{\mathrm{roll}}$ index the online subset, of size $m_s \in \{G, M\}$. Only rollouts are scored during training. For $R_i = R(c_s, x_1^{(i)})$ and $\gamma > 0$, define

$$\phi_s = \frac{1}{m_s} \sum_{k \in \mathcal{I}_s^{\mathrm{roll}}} \exp(R_k/\gamma), \qquad a_s^{(i)} = \begin{cases} \widehat{\mathcal{A}}_{0,s}^{(i)} = \exp(R_i/\gamma)/\phi_s - 1, & i \in \mathcal{I}_s^{\mathrm{roll}}, \\ 1, & \text{otherwise.} \end{cases} \tag{15}$$

Exemplars do not enter ϕ_s; rollout advantages sum to zero within their subset. A single rollout would therefore have zero centered advantage. Write $v_\theta^{(i)} = v_\theta(x_{t_s}^{(i)}, c_s, t_s)$, likewise for $v^{\mathrm{old},(i)}$, and $u_{t_s}^{(i)} = x_1^{(i)} - x_0^{(i)}$. The group-averaged loss is

$$\widehat{\mathcal{L}}_{\mathrm{GraftOOD}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \left\| v_\theta^{(i)} - \left[v^{\mathrm{old},(i)} + a_s^{(i)} \left(u_{t_s}^{(i)} - v^{\mathrm{old},(i)} \right) \right] \right\|_2^2. \tag{16}$$

For an exemplar, $a_s^{(i)} = 1$ cancels the reference field pointwise, so using v^{old} here gives the same target u_t as the source-specific expression in Equation 12. Targets are detached. For an exemplar prompt, the expected group average is Equation 12; original prompts retain only the online term.

For a fixed temperature, independent re-noising makes x_0 independent of the terminal given the prompt. The group normalizer estimates the noise-end denominator $\phi(x_0, c, 0)$ from Equation 6. Reusing $\widehat{\mathcal{A}}_0$ at other timesteps approximates, rather than computes, their state-conditional coefficients. The subscript zero denotes conditioning time; rewards are evaluated on clean terminals.

A.2 SAMPLING ASSUMPTIONS AND PROOF

Sampling model and assumptions. Fix c_s, the reference parameters, and the group-sampling law for one optimization round. Draw a group with M online rollouts and n_{ood} exemplars, then select one member uniformly. Its source $S \in \{\mathrm{on}, \mathrm{ood}\}$ has prior $\Pr(S = \mathrm{ood} \mid c_s) = \delta = n_{\mathrm{ood}}/G$. The source-conditional terminal laws are $p^{\mathrm{old}}(\cdot \mid c_s)$ and $p^{\mathrm{ood}}(\cdot \mid c_s)$, with the latter reflecting the exemplar selection rule. Draw the shared $t_s \sim \nu$ independently of the collected group and re-noise each member independently. Linearity of expectation makes this shared draw equivalent in expected loss to independent per-member draws from ν, though their variances may differ. All almost-everywhere statements below refer to this sampling law, including when ν is discrete.

For the selected member, set $Z = (c_s, x_t, t)$ with $t = t_s$ and $Y = v^{\text{old}}(Z) + a(u_t - v^{\text{old}}(Z))$, using Equation 15. Assume $\mathbb{E}\|Y\|_2^2 < \infty$ and that the two noised source laws admit densities with respect to a common dominating measure for ν-almost every t. Uniform member selection reproduces Equation 12. Wherever the mixture density is positive, Bayes' rule gives $\Pr(S = \text{ood} \mid Z) = w_s(x_t, t)$ in Equation 14. Conditional means need only be defined on each source's support; a zero-posterior term is omitted.

Proof of Proposition 1. Under the fixed-prompt sampling law above, the conditional variance decomposition gives, for any measurable field v of finite risk,

$$\mathbb{E}\|v(Z) - Y\|_2^2 = \mathbb{E}\|v(Z) - \mathbb{E}[Y \mid Z]\|_2^2 + \mathbb{E}\|Y - \mathbb{E}[Y \mid Z]\|_2^2. \tag{17}$$

The second term is independent of v, so the minimizer is $v^\dagger(Z) = \mathbb{E}[Y \mid Z]$, unique up to sets of training-state measure zero. Define $v^\star_{\text{old}}(Z) = \mathbb{E}[Y \mid Z, S = \text{on}]$. For exemplars, $Y = u_t$ and conditional flow matching gives $\mathbb{E}[Y \mid Z, S = \text{ood}] = v^{\text{ood}}(Z)$. Conditioning on the source therefore yields

$$v^\dagger(Z) = (1 - w_s)v^\star_{\text{old}}(Z) + w_s v^{\text{ood}}(Z),$$

which is Equation 13. The online conditional mean includes the other rollout members that determine the finite-group coefficient and the rollout-batch statistics used to set γ; no population-advantage substitution is needed.

This is a characterization over measurable fields. The empirical online target need not equal the exact KL-optimal field, and zero-sum group advantages do not cancel reference fields evaluated at different states. The result does not establish neural-network attainability, training convergence, terminal-distribution mixing, or video-quality improvement.

B Experimental Details and Additional Results

B.1 GraftSet-Video Details

Source counts and selection. Table 2 gives exact source counts, supplementing the proportions and mean HPSv3 scores in Figure 3. Construction and training-time allocation follow Section 3.3: supplementary outputs fill unoccupied exemplar slots rather than replacing retained videos through a pooled reranking of collected and supplementary candidates.

Corpus statistics. HPSv3 scores span $[-0.139, 15.131]$, with mean 9.735 and standard deviation 1.413. Of the exemplar prompts, 52.0% contain Chinese text, and the mean prompt length is 388 characters. Video format and prompt-set separation are specified in Section 3.3.

Table 2: Exemplars by source.

System	Videos
`wan2_2`	2,164
`vidu_q2`	1,163
`seedance_1_5`	1,144
`kling_2_5`	1,098
`kling3`	915
`seedance2`	911
`veo3_1`	883
`hailuo`	830
`sora2`	480
Total	9,588

B.2 Experimental Setup

Training and sampling configurations. All experiments use Wan2.1-1.3B (Wan et al., 2025) and 19,700 training prompts following DanceGRPO (Xue et al., 2025), with additional exemplar prompts and videos from GraftSet-Video (Section 3.3). Videos contain 81 frames at 832×480 resolution and 16 frames per second. Online sampling uses classifier-free guidance (CFG); NFT/AWR variants disable CFG for training updates. All reported evaluations, including the base model, use 50 sampling steps and a CFG scale of 4.5. Training rollouts use 20 sampling steps, a CFG scale of 4.5, and group size $G = 8$; NFT/AWR variants use EMA reference policies for sampling. We train rank-32 LoRA adapters with scaling factor 64 on the attention projections, keeping the backbone frozen. We use bfloat16 and AdamW with learning rate 10^{-4}, weight decay 10^{-4}, and gradient norm clipping at 1.0. FlowAWR and GraftOOD set γ to the global standard deviation of online-rollout rewards in the current batch, shared across prompt groups and excluding exemplars. FlashOOD uses

$\lambda_{\text{ood}} = 1$ in all configurations. GraftOOD shares a regression timestep within each prompt group and independently re-noises each video.

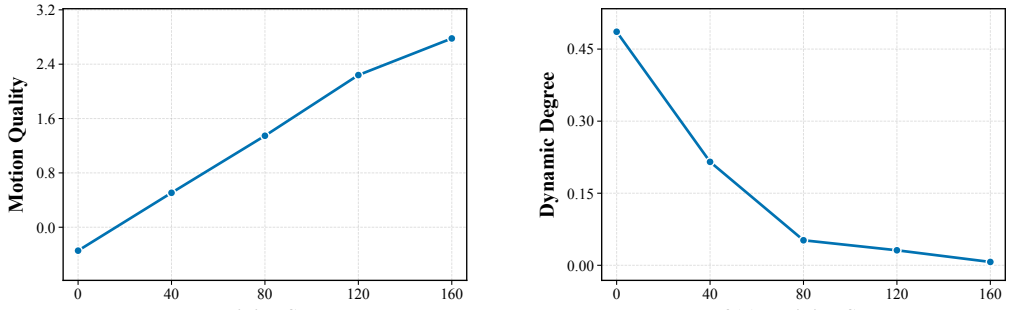
The main comparison uses $\rho = 0.167$ and $n_{\text{ood}} = 4$. The ablations separate two controls: the frequency of exemplar prompt groups, ρ, and their within-group exemplar count, n_{ood}. The experimental grid uses $\rho \in \{0.167, 0.33, 0.5\}$ and $n_{\text{ood}} \in \{4, 2, 1\}$ with $G = 8$, retaining four, six, or seven online rollouts per exemplar group, respectively. All configurations use the same GraftSet-Video candidate pool, with four stored videos per prompt. Reducing n_{ood} changes how many candidates are used in a group, not the corpus or prompt population.

Reward and evaluation protocols. The main experiments optimize HPSv3 (Ma et al., 2025), averaging the top 30% of per-frame scores following LongCat-Video (Team et al., 2025b). For offline corpus construction, the initial selection from public resources retains videos with HPSv3 scores above 7, before supplementary generation fills the remaining exemplar slots. This initial threshold does not impose a score floor on the supplemented corpus. External exemplars do not enter training-time reward normalization. We report VBench's Aesthetic Quality (AQ), Imaging Quality (IQ), Subject Consistency (SC), and Object Class (OC) on its official prompt set (Huang et al., 2024). On the 300-prompt VidProM evaluation set, we measure Dynamic Degree (DD), VisionReward (VR) (Xu et al., 2026), VideoAlign's Video Quality (VQ), Motion Quality (MQ), and Text Alignment (TA) (Liu et al., 2026b), and HPSv3. Motion Quality serves as an evaluation metric here; its use as a training reward in the motivation study is discussed in Section 3.1.

These evaluation sets remain unchanged across prompt ratios and exemplar counts. DD measures detected dynamics rather than motion plausibility and should be read alongside the other metrics and qualitative examples in Section 4.2.

B.3 Additional Quantitative Results

Motion-quality reward optimization. Figure 7 supplements the motivation study in Section 3.1. During optimization of VideoAlign Motion Quality, its reward increases while Dynamic Degree decreases. This example shows that using a video-level motion-quality reward does not by itself prevent the loss of measured dynamics.



(a) Training step (b) Training step

Figure 7: **Motion-quality reward and video dynamics.** VideoAlign Motion Quality reward (left) and Dynamic Degree (right) during the same reward-optimization run.

Method comparisons at four exemplars. Tables 3–5 compare FlashOOD, NFTOOD, and GraftOOD at each ratio with $n_{\text{ood}} = 4$. Bold values denote the best result among these three methods in the same column and table, not across ratios. The $\rho = 0.167$ results repeat the OOD rows of Table 1; the three tables provide the positive-ratio points in Figure 4. The $\rho = 0$ points in that figure instead use each method's reward-only baseline from the main table. GraftOOD has the highest HPSv3 and VR at all three positive ratios, while FlashOOD has the highest DD. Within GraftOOD, increasing ρ raises DD but lowers HPSv3 and VR over the reported range.

Table 3: OOD comparison at $\rho = 0.167$, $n_{\text{ood}} = 4$. Bold marks column bests.

Method	AQ	IQ	SC	OC	DD	VR	VQ	MQ	TA	HPSv3
FlashOOD	65.87	67.91	94.68	88.29	**58.26**	1.65	-0.62	-0.47	-0.50	5.31
NFTOOD	66.10	70.52	96.99	87.91	41.12	3.90	-0.45	-0.25	-0.47	6.55
GraftOOD	**67.21**	**71.26**	**98.20**	**90.10**	44.41	**5.08**	**-0.25**	**-0.07**	**-0.45**	**7.32**

Table 4: OOD comparison at $\rho = 0.33$, $n_{\text{ood}} = 4$. Bold marks column bests.

Method	AQ	IQ	SC	OC	DD	VR	VQ	MQ	TA	HPSv3
FlashOOD	65.78	67.17	95.05	88.21	**51.60**	1.87	-0.52	-0.38	-0.49	5.62
NFTOOD	**66.41**	69.11	96.95	89.85	43.12	4.04	-0.38	-0.21	**-0.46**	6.61
GraftOOD	66.34	**71.39**	**97.04**	**90.19**	45.72	**4.76**	**-0.32**	**-0.14**	-0.50	**6.94**

Table 5: OOD comparison at $\rho = 0.5$, $n_{\text{ood}} = 4$. Bold marks column bests.

Method	AQ	IQ	SC	OC	DD	VR	VQ	MQ	TA	HPSv3
FlashOOD	66.10	68.40	94.72	87.10	**55.13**	1.96	-0.63	-0.49	-0.51	5.68
NFTOOD	**66.27**	69.01	95.77	**89.88**	45.36	3.52	**-0.37**	**-0.14**	-0.46	6.21
GraftOOD	66.03	**69.21**	**96.75**	88.92	48.68	**3.97**	-0.43	-0.25	**-0.39**	**6.68**

GraftOOD ratio and count ablations. Tables 6–8 arrange the GraftOOD results by exemplar count. All three tables include the same $\rho = 0$ FlowAWR baseline from Table 1. At zero ratio, no exemplar groups are sampled and n_{ood} is inactive; these entries repeat one baseline rather than represent separate count-specific runs. Reading down a table varies the frequency of exemplar groups while keeping their composition fixed at positive ratios. Comparing the same positive ρ across tables varies the number of exemplars within each such group while keeping its sampling frequency fixed. The positive-ratio rows of Table 6 collect the GraftOOD results from Tables 3–5; they do not represent additional runs. Tables 7 and 8 report the corresponding results for smaller counts. Figure 6 compares HPSv3, DD, and VisionReward across these settings. Bold values mark column bests among positive ratios within each fixed-count table; the shared $\rho = 0$ baseline is excluded.

Under the group sampling scheme in Section 3.3, the fraction of training samples supplied by exemplars is $\rho n_{\text{ood}}/8$. Thus, increasing either control increases exemplar exposure, but the two interventions differ: ρ changes the frequency of exemplar prompts, whereas n_{ood} also changes the number of online samples within those groups. The latter affects the sample set used for online reward normalization. The two comparisons therefore assess distinct training configurations rather than interchangeable settings of a single loss weight.

Table 6: GraftOOD ratio sweep with $n_{\text{ood}} = 4$ ($G = 8$). The $\rho = 0$ row repeats FlowAWR; positive rows repeat Tables 3–5. Bold marks column bests among positive ratios only.

ρ	AQ	IQ	SC	OC	DD	VR	VQ	MQ	TA	HPSv3
0	69.18	71.01	98.48	92.53	4.86	5.11	0.06	0.47	-0.65	8.10
0.167	**67.21**	71.26	**98.20**	90.10	44.41	**5.08**	**-0.25**	**-0.07**	-0.45	**7.32**
0.33	66.34	**71.39**	97.04	**90.19**	45.72	4.76	-0.32	-0.14	-0.50	6.94
0.5	66.03	69.21	96.75	88.92	**48.68**	3.97	-0.43	-0.25	**-0.39**	6.68

Table 7: GraftOOD ratio sweep with $n_{\text{ood}} = 2$ ($G = 8$). The $\rho = 0$ row is the shared FlowAWR baseline. Bold marks column bests among positive ratios only.

ρ	AQ	IQ	SC	OC	DD	VR	VQ	MQ	TA	HPSv3
0	69.18	71.01	98.48	92.53	4.86	5.11	0.06	0.47	-0.65	8.10
0.167	66.59	**71.29**	98.18	**90.59**	31.41	**5.34**	**-0.03**	**0.15**	-0.53	**7.6**
0.33	**67.57**	71.08	**98.25**	88.26	40.79	4.89	-0.17	0.09	-0.5	7.06
0.5	66.25	69.43	97.17	89.71	**45.72**	3.9	-0.47	-0.28	**-0.47**	6.54

Table 8: GraftOOD ratio sweep with $n_{\text{ood}} = 1$ ($G = 8$). The $\rho = 0$ row is the shared FlowAWR baseline. Bold marks column bests among positive ratios only.

ρ	AQ	IQ	SC	OC	DD	VR	VQ	MQ	TA	HPSv3
0	69.18	71.01	98.48	92.53	4.86	5.11	0.06	0.47	-0.65	8.10
0.167	**68.21**	**72.10**	98.17	88.38	30.58	**5.52**	-0.32	-0.13	**-0.39**	**7.74**
0.33	67.43	71.23	**98.27**	90.66	33.93	5.09	**-0.13**	**0.05**	-0.57	7.36
0.5	66.81	69.31	98.01	**91.48**	**46.76**	4.69	-0.38	-0.17	-0.43	7.13

C Limitations and Future Work

Evaluation scope and exemplar sources. Our experiments are limited to text-to-video generation with Wan2.1-1.3B, and all exemplars in GraftSet-Video are generated videos. The $n_{\text{ood}} = 1$ ablation evaluates a single-exemplar group configuration, but does not establish the effectiveness of real-video supervision. The current results therefore leave generalization to other tasks, model scales, and exemplar sources untested.

Real-video supervision and physical consistency. Real-world recordings could provide complementary examples of object interactions and temporal behavior that generated exemplars may reproduce imperfectly. Whether such supervision improves physical consistency remains an empirical question, requiring suitable video–text pairing, data selection, and evaluation. GraftOOD offers a regression interface for investigating this direction because exemplar supervision requires terminal videos rather than their generating denoising trajectories; this does not by itself guarantee successful transfer to real-video data.

Extension to audio–video generation. A further direction is text-to-audio-video generation (T2AV), where supervision must account for both visual dynamics and audio–visual correspondence. Such an extension would require appropriate joint representations and reward models sensitive to temporal synchronization and cross-modal consistency. More broadly, advances in rewards and evaluation remain necessary to determine whether exemplar-guided optimization improves physical plausibility rather than merely increasing proxy scores.



Figure 8: **Additional qualitative comparisons.** Each block shows Vanilla, FlashOOD, NFTOOD, and GraftOOD from top to bottom, with three frames in temporal order. Full prompts appear in Appendix D.2.

D More Qualitative Evaluation

D.1 Additional Comparisons

We extend the qualitative comparisons in Figures 1 and 5 with additional examples in Figure 8. We compare the vanilla generator with FlashOOD, NFTOOD, and GraftOOD. Each example presents three representative frames in temporal order, allowing a side-by-side inspection of subject appearance, scene composition, and visible changes across the sampled frames.

The additional examples cover both natural scenery and stylized animal characters. In the sunset scene, the GraftOOD example combines a prominent sailboat with textured clouds, warm illumination, and reflections on the water. In the trumpet scene, it depicts a richly colored stage and audience behind the performing raccoon. The bear-and-zebra example illustrates differences in character rendering and meadow detail, while the polar-bear example highlights fur, guitar, and snow-covered forest textures. Across these selected examples, GraftOOD produces visually detailed scenes with recognizable subjects and scene layouts across the displayed frames. These qualitative observations complement the quantitative evaluation.

D.2 Full Prompts for the Qualitative Comparisons

Prompts in Figure 1. The examples are listed in the order top to bottom.

1. a cat is running on the snow ground.
2. a man standing on top of the mountain,a scarf and hair in wind, move dolly camera to left.
3. A boat sailing leisurely along the Seine River with the Eiffel Tower in background by Vincent van Gogh.

4. 20 years old anime girl with long dark hair nice body sitting in the woods and the hair is waving with the wind, windy, lighting, warm light, sun, sun rays.

Prompts in Figure 5. The examples are listed in the order top-left, top-right, bottom-left, and bottom-right.

1. In late autumn inside a greenhouse, a gardener in a soft cardigan and floral apron fills the foreground in cinematic 4K. She trims a rose, then mists the leaves as the camera moves close to droplets hanging from petals. Outside the glass, rain falls and trees turn bronze, while inside warm light makes everything glow. Crisp textures, true tones, and gentle film grain create a cozy, intimate beauty.
2. A scenic photograph capturing the moment a steam train departs from the Glenfinnan Viaduct, a historic railway bridge in Scotland. The train moves gracefully over the arch-covered viaduct, its smoke billowing into the air. The landscape is lush with greenery, and towering rocky mountains frame the scene, creating a picturesque backdrop. The sky is a clear, bright blue with the sun shining down, casting a warm glow on the train and the surrounding scenery. The viaduct itself is a striking feature, with intricate ironwork and a verdant setting. The photo has a classic, nostalgic feel, emphasizing the natural beauty and historical charm of the location. A wide-angle shot from a slightly elevated angle, capturing both the train and the expansive landscape.
3. A cute raccoon playing guitar in a boat on the ocean.
4. A green dinosaur and a yellow kitten are playing basketball on a football field.

Prompts in Figure 8. The examples are listed in the order top-left, top-right, bottom-left, and bottom-right.

1. Crear un velero de 50 pies hiper realista navegando hacia la pusta de sol.
2. A playful raccoon dressed in a stylish black suit with a bright red bow tie, playing a shiny silver trumpet on a bustling city stage. The raccoon has expressive brown eyes and a charming smile, standing confidently in front of a large, colorful backdrop featuring vibrant lights, a busy crowd, and a neon sign advertising the concert. The raccoon's fur is sleek and glossy, highlighting its playful nature as it takes center stage, surrounded by enthusiastic audience members. The stage background includes intricate patterns and detailed architectural elements, creating a lively and dynamic atmosphere. Close-up, medium shot, and wide shot angles capturing the raccoon's performance from various perspectives. Smooth, vibrant animation style.
3. A whimsical wildlife scene captured in a playful animated style, featuring a majestic brown bear and a vibrant black and white zebra standing side-by-side. The bear stands tall with a gentle smile, its fur soft and fluffy, while the zebra looks alert and curious with its striped mane blowing in the breeze. They share a moment of quiet companionship in a lush green meadow dotted with wildflowers. The sky above is a vivid shade of azure with fluffy cotton candy clouds. The background showcases a serene landscape with rolling hills and a babbling brook nearby. The entire scene is rendered in a lively hand-drawn animation style with bright, cheerful colors. Close-up shot of the two animals facing each other, medium shot of their interaction, and a panoramic view of the meadow.
4. CG fantasy digital art, a majestic polar bear with a fluffy white coat and large black patches on its face, standing on its hind legs, holding a small acoustic guitar with a wooden body and intricate carvings. The bear's fur is meticulously detailed, with soft highlights and shadows. It is surrounded by a snow-covered forest with tall pine trees and sparkling ice formations. The bear is playing a lively melody, its eyes focused intently on the strings. The atmosphere is magical and enchanting. Low-angle view, focusing on the bear's expressive face and the intricate details of the guitar.