EMOREWARD: SELF-EVOLVING REWARD CALIBRATION FOR EMOTIONAL VIDEO GENERATION RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Relative Policy Optimization (GRPO) has advanced generative model post-training in verifiable domains such as code and mathematics, but evaluating video quality, especially emotional expression, often depends on subjective aesthetics and has no uniquely correct target. Existing approaches typically rely on fixed rewards for visual quality, motion quality, and text alignment that poorly capture emotional intent and can invite reward hacking, while reinforcement learning from human feedback (RLHF) is difficult to scale and subject to individual aesthetic differences that complicate the calibration of facial, body motion, and chromatic cues across narrative contexts. We introduce **EmoReward**, a new self-evolving training mechanism that distills evidence reliability from reward traces collected during an actual training run to adaptively guide multi-reward weight allocation in subsequent training by determining which emotional evidence to trust and how much to weight it. EmoReward first plans specific grounding cues from the emotional prompt to guide evaluation, then uses heterogeneous signal primitives to score the generated video during the act stage. *Reflective Reward Gating* calibrates their contributions through instance-level gates for phase, face confidence, conflict, and subtlety, then renormalizes weights over the active set to produce the sample's fused reward. *RewardMemory* then estimates each primitive's reliability from rollout scores and outcomes within each emotional regime in the finished training run, then fixes bounded multipliers that calibrate fusion weights in the next training run without updating reward models. Experiments on HunyuanVideo 1.5 show improvements across multiple RL algorithms and compare EmoReward with other agentic reward methods under SAGE-GRPO. Ablations and transfer experiments on SD 3.5 and FLUX further support adaptive reward calibration for visual generation.

1 INTRODUCTION

Reinforcement learning (RL), especially Group Relative Policy Optimization (GRPO), has advanced model post-training in verifiable domains such as code and mathematics, where correctness provides a direct learning signal (Shao et al., 2024; Guo et al., 2025; Kimi Team, 2025; Yu et al., 2025). In contrast, evaluating video quality, especially emotional expression, often depends on subjective aesthetics and has no uniquely correct target. Visual post-training adapts diffusion and flow generators through RL and preference objectives (Black et al., 2023; Fan et al., 2023; Prabhudesai et al., 2023; Wallace et al., 2023; Xue et al., 2025; Liu et al., 2025a; Zheng et al., 2026), with recent work also exploring on-policy distillation (Jiang et al., 2026; Fang et al., 2026). Learned evaluators provide scores for image preference and video quality (Xu et al., 2023; Kirstain et al., 2023; Ma et al., 2025; Xu et al., 2024; He et al., 2024; Liu et al., 2025b), but a visually clear, temporally stable video can match its prompt without conveying the intended emotion. Facial analysis, dynamic expression recognition, and multimodal emotion models supply complementary evidence (Baltrušaitis et al., 2018; Chen et al., 2024; Cheng et al., 2024; Lian et al., 2025; Zhao et al., 2025), whose reliability varies with context: facial cues require visible faces, strong motion can undermine restrained sadness, and saturated colors may reflect style rather than emotional intent. This raises a central question: *Which emotional evidence should the reward system trust in the current video, and how can it adaptively determine how much weight to assign to that evidence?*

Agentic systems coordinate interaction (Wu et al., 2023a; Wang et al., 2025), refine outputs through feedback (Madaan et al., 2023; Shinn et al., 2023), and organize evaluation into rubrics (Gunjal et al.,

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.

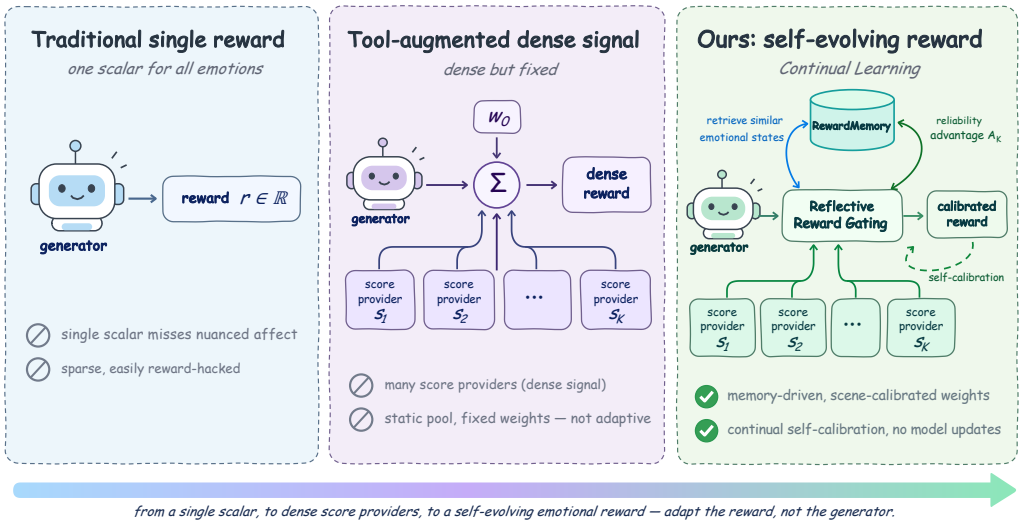


Figure 1: **From fixed scoring to contextual reward calibration.** The reward design progresses from a fixed evaluator to a heterogeneous provider pool and then to EmoReward. Reflective Reward Gating selects relevant evidence for the current video; RewardMemory calibrates provider trust from experience. All three stages ultimately return a scalar reward. The reward providers remain frozen while the generator undergoes RL updates.

2025), with recent extensions to image generation and multimodal search (Chen et al., 2026b; Feng et al., 2026; Chen et al., 2026a). These capabilities motivate adaptive evaluation, yet existing video GRPO methods typically rely on fixed rewards for visual quality, motion quality, and text alignment that inadequately capture emotional intent. Reinforcement learning from human feedback (RLHF) and preference optimization address subjective goals (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023), but human annotation is costly at scale, and aesthetic differences complicate the calibration of emotional cues across contexts. Fixed fusion can mix informative, missing, and saturated signals, allowing policy updates to exploit scoring shortcuts rather than improve emotional fidelity. This motivates an adaptive agentic reward mechanism that assesses which cues are valid for the current video and uses accumulated outcomes to calibrate their relative influence, reducing reliance on fixed scores vulnerable to reward hacking.

We introduce **EmoReward**, a self-evolving reward calibration framework with frozen reward providers (Figure 1). *Plan* translates emotional intent into an EmoScript specifying affect targets, facial and body cues, and an expression window. *Act* evaluates generated videos with heterogeneous primitives, keeping criterion scores separate from evidence validity. *Reflective Reward Gating* adjusts their contributions using training phase, face confidence, semantic conflict, and subtlety, then renormalizes over active providers. *RewardMemory* retains reward–outcome traces from a completed reference run, estimates provider reliability within emotional regimes, and converts centered reliability into bounded fusion multipliers. These multipliers remain fixed during the subsequent run and combine with current-video gates: memory supplies an experience-based allocation, while reflection limits it to evidence valid for the current scene. The calibrated emotion reward complements the original video reward, and only the generator receives gradient updates. Thus, one completed training process becomes reusable guidance for the next without retraining the evaluators.

We evaluate EmoReward on HunyuanVideo 1.5 with DanceGRPO(Xue et al., 2025), FlowGRPO(Liu et al., 2025a), and SAGE-GRPO(Zheng et al., 2026), observing higher VideoAlign scores for the guided configurations across all three optimizers on 302 emotional prompts as the evaluation dataset. Under a shared SAGE-GRPO backbone and training budget, EmoReward achieves the highest reliability selectivity and both composite calibration scores among seven reward policies. Ablation studies and mechanism analyses examine the contributions of reflective gating, provider pools, and RewardMemory (Section 4.2). Transfer experiments on Stable Diffusion 3.5 and FLUX show gains

in selected configurations alongside objective-dependent trade-offs, while user studies show that EmoReward receives the highest human preference and proportion of Good ratings and the lowest proportion of Bad ratings for emotional fidelity, temporal coherence, and visual quality.

Our contributions are threefold.

- We introduce **EmoReward**, to our knowledge the first self-evolving reward evaluation framework for emotional video generation RL that calibrates reward providers across training runs without updating reward model parameters.
- We propose **RewardMemory**, which treats a completed training run as a source of reusable reward–outcome traces that guide subsequent post-training. This provides a practical approach to reward-side memory across training runs, beyond experience reuse within a single run.
- We evaluate EmoReward across three video RL algorithms and two image backbones, supported by ablations and human ratings. Under a shared SAGE-GRPO backbone and optimization budget, EmoReward outperforms the compared agentic reward policies on both composite calibration metrics; transfer experiments characterize its applicability and objective-dependent trade-offs.

2 Related Work

From verifiable RL to agentic feedback for non-verifiable rewards. RLHF learns subjective preferences (Christiano et al., 2017; Ouyang et al., 2022), whereas verifiable RL exploits correctness checks (Guo et al., 2025; Yu et al., 2025). Agentic evaluators (Zhang et al., 2025; Cui et al., 2026) and adaptive rubrics (Gunjal et al., 2025; Xu et al., 2026; Zhang et al., 2026) structure open-ended feedback. Related approaches route experts (Ong et al., 2024), refine responses (Madaan et al., 2023), evolve rewards (Ma et al., 2024; Yuan et al., 2024), or organize experience (Shinn et al., 2023; Xu et al., 2025). JitRL derives closed-form policy updates from retrieved action advantages (Li et al., 2026). EmoReward instead transfers reward–outcome evidence across completed training runs to calibrate frozen providers, combining sample-level gating with persistent fusion weights.

RLHF for visual generation and the emotion-reward gap. Visual reward models capture preferences and multidimensional quality (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023b; Ma et al., 2025; Xu et al., 2024; Wang et al., 2026b; Liu et al., 2025b). Optimization using rewards and preferences extends from images (Black et al., 2023; Fan et al., 2023; Prabhudesai et al., 2023; Wallace et al., 2023) to video and flow models (Xue et al., 2025; Liu et al., 2025a; Zheng et al., 2026; Liu et al., 2025c), with benchmarks assessing quality and compositional alignment (Huang et al., 2024; Liu et al., 2024; Sun et al., 2025). Related 3D, audiovisual, and objectives for human animation address complementary aspects of visual generation (Leng et al., 2026; Huang et al., 2026; Tu et al., 2025b;a). Affect-grounded augmentation (Wang et al., 2026a) and reflective action-conditioned generation (Yang et al., 2026) further connect feedback with video synthesis. Emotion understanding supplies multimodal evidence (Cheng et al., 2024; Lian et al., 2025; Zhao et al., 2025) and grounded affect representations (Russell, 1980; Ekman & Friesen, 1978; Baltrušaitis et al., 2018); EmoReward addresses how their reliability should govern reward fusion across scenes and training runs.

3 Method

3.1 Preliminaries

Multi-reward optimization with GRPO. Let p be a prompt and x a generated video from a policy $\pi_\theta(x \mid p)$; a scalar reward is $R(x, p) \in \mathbb{R}$. For a prompt p, GRPO samples a group $\mathcal{G}(p) = \{x_i\}_{i=1}^{G}$ and forms standardized group advantages

$$A_i = \frac{R(x_i, p) - \mu_{\mathcal{G}}}{\sigma_{\mathcal{G}} + \epsilon}, \qquad \mu_{\mathcal{G}} = \tfrac{1}{G}\textstyle\sum_j R(x_j, p), \quad \sigma^2_{\mathcal{G}} = \tfrac{1}{G}\textstyle\sum_j (R(x_j, p) - \mu_{\mathcal{G}})^2, \tag{1}$$

which enter the clipped, KL-regularized objective in Appendix F. The policy update depends on reward differences inside the group. A common additive constant cancels exactly; positive rescaling

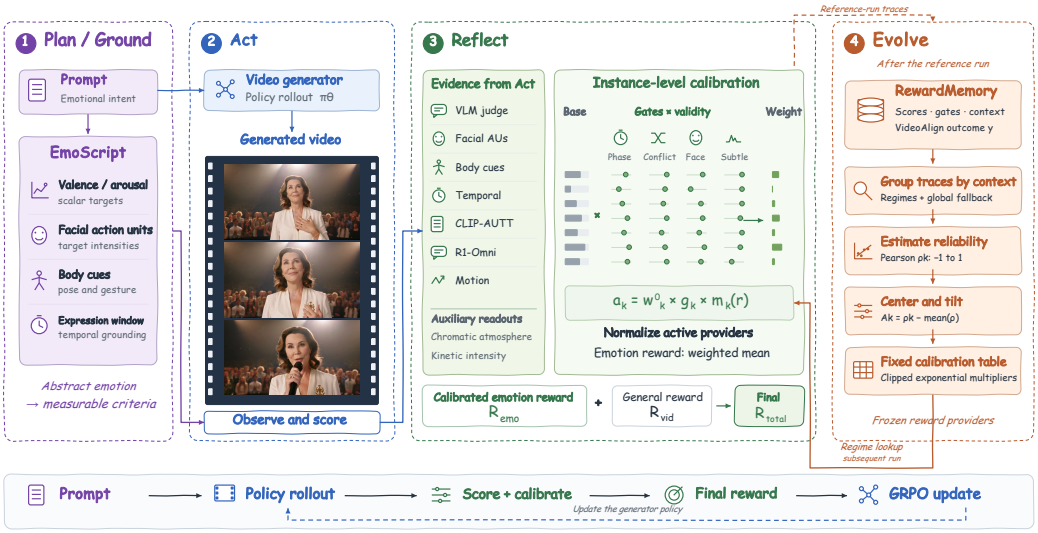


Figure 2: **EmoReward calibrates emotional evidence within videos and across training runs.** Plan translates intent into evaluation criteria, and Act obtains provider scores and evidence validity. Reflect combines current-video gates with memory-derived multipliers and normalizes over active providers. Evolve estimates contextual reliability from a completed reference run to construct the fixed multipliers used in subsequent training. The calibrated emotion reward complements the original video reward; GRPO updates the generator while the providers and memory table remain frozen.

changes only the relative effect of ϵ. A coefficient of variation measures dispersion, not policy-learning signal quality; we also assess within-prompt ranking agreement and downstream quality.

Why static emotion fusion fails. Given reward primitives $\mathcal{T} = \{1, \ldots, K\}$ with normalized scores $s_k(x, p) \in [0, 1]$, a fixed fusion is $R_{\text{emo}}^{\text{static}} = \sum_k w_k^0 s_k$ with $w_k^0 \geq 0, \sum_k w_k^0 = 1$. This assumes a fixed trust allocation across scenes, although facial visibility and the relevance of motion depend on the intended emotion. Writing each primitive as a constant plus a group-varying part $s_k = c_k + \delta_k$ and splitting into an active set $\mathcal{A}$ and a dead/saturated set $\mathcal{D}$,

$$R_{\text{emo}}^{\text{static}} = \sum_{k \in \mathcal{A}} w_k^0 \delta_k + \sum_{k \in \mathcal{A}} w_k^0 c_k + \sum_{k \in \mathcal{D}} w_k^0 c_k, \tag{2}$$

the constant terms cancel from the centered GRPO numerator. Removing them alone need not change the standardized advantage. Calibration matters when provider validity, errors, or relative contributions vary across candidates, or when the emotion branch is mixed with another reward. EmoReward addresses these cases through gates and memory-conditioned fusion.

3.2 Overview

EmoReward grounds emotional intent, scores visual evidence, and calibrates provider trust:

$$p \xrightarrow{\text{Plan}} e(p), \quad (x, e) \xrightarrow{\text{Act}} \{s_k, u_k\}, \quad (s, u, e, x) \xrightarrow{\text{Reflect}} R_{\text{emo}}, \quad \mathcal{M} \xrightarrow{\text{Evolve}} w^{\text{RM}}(z). \tag{3}$$

Here $e(p)$ is an EmoScript, u_k denotes evidence validity, and z is the emotional-context signature. RewardMemory $\mathcal{M}$ stores completed reference-run traces; Evolve produces $w^{\text{RM}}(z)$ and exports the corresponding regime multipliers $m_k(r)$ (Section 3.5). Reflect combines these fixed multipliers with current-video gates. The resulting emotion reward is mixed with a general video reward:

$$R_{\text{total}}(x, p) = \frac{\alpha R_{\text{vid}}(x, p) + \beta R_{\text{emo}}(x, p)}{\alpha + \beta}, \tag{4}$$

where R_{vid} is the evaluation score supplied by the original reward system, for example a general video-quality evaluator (Liu et al., 2025b); $\alpha, \beta \geq 0$ control its mixture with the calibrated emotion reward, with $\alpha + \beta > 0$. Figure 2 connects these stages through their shared fusion interface.

Algorithm 1 EmoReward reward computation with experience-calibrated weights

Require: prompt p, generated video x, fixed table $\mathcal{T}_{\mathcal{M}}$ from a completed reference run, base weights w^0

Build EmoScript $e(p)$
Compute primitive scores and validity $\{s_k, u_k\}_{k=1}^{K}$ ▷ Eq. 6
Compute instance gates $\{g_k\}_{k=1}^{K}$ ▷ Eq. 7
Form state signature $z \leftarrow z(p, x, e, s, g)$
Look up regime $r(z)$ in $\mathcal{T}_{\mathcal{M}}$ ▷ Appendix E
if a supported regime or global entry is available **then**
    Read fixed multipliers m_k, constructed by Eq. 10
else
    $m_k \leftarrow 1$ ▷ cold start: fall back to base
end if
Form active weights $a_k \leftarrow w_k^0 m_k g_k$
Fuse active primitives with renormalization $\rightarrow R_{\text{emo}}$ ▷ Eq. 8
Compute total reward $R_{\text{total}}(x, p)$ ▷ Eq. 4
Log the reward trace for inspection; keep $\mathcal{T}_{\mathcal{M}}$ fixed
return $R_{\text{total}}(x, p)$

3.3 Plan and Act: Emotional Grounding

Plan: prompt-conditioned criteria. The planner converts the prompt into an EmoScript independently of RewardMemory:

$$e(p) = (v^\star, a^\star, \mathbf{q}^\star_{\text{AU}}, \mathbf{b}^\star, \Omega). \tag{5}$$

The implemented representation contains scalar valence and arousal targets, target facial AU intensities, body cues, and an expression window. The memory regime is inferred later from observed context. Continuous VAD trajectories and dominance targets are outside this implementation. The EmoScript separates intended affect from its observable cues. For example, restrained sadness need not imply large motion or an exaggerated face: target intensities, body cues, and the expression window define what to assess, while reliability is determined from the generated evidence and completed-run experience.

Act: observation and scoring. Act evaluates a video already sampled from $\pi_\theta(\cdot \mid p)$; it does not generate or revise the video. Each provider returns

$$(s_k, u_k) = \mathcal{S}_k(x, e(p)), \qquad s_k, u_k \in [0, 1], \tag{6}$$

where $\mathcal{S}_k$ is provider k's scoring operator, which aggregates frame-level measurements or evaluates the whole clip. The seven fusion providers comprise a VLM judge, facial AUs, body cues, temporal evidence, CLIP-AUTT, R1-Omni, and motion naturalness; chromatic atmosphere and kinetic intensity are auxiliary readouts, giving nine observation channels. Score s_k measures criterion agreement, while u_k marks usable evidence. A missing face channel's neutral default is not evidence and is excluded through its validity mask. Auxiliary readouts do not enter the seven-provider calibration policy. Appendix C specifies scoring interfaces; Figure 5 illustrates the toolkit studied in Section 4.2.

3.4 Reflect: Instance-Level Calibration

Reflect computes provider-specific gates from the current evidence, prompt, and training step. The factorization is

$$g_k = g_k^{\text{phase}}\, g_k^{\text{conflict}}\, g_k^{\text{face}}\, g_k^{\text{subtle}}\, u_k, \qquad g_k \in [0, 1], \tag{7}$$

with neutral factors for channels to which a condition does not apply. Phase denotes the training-step warmup, face gating controls face-dependent evidence, conflict gating attenuates disagreement between semantic providers, and subtlety gating restrains intensity-sensitive cues. Motion naturalness uses its observation confidence. These controls are computed, not learned (Appendix D).

Evidence-dependent trust. The phase factor schedules reward evidence over generator training, distinct from the within-video expression window. Face gating limits channels requiring facial

Algorithm 2 RewardMemory construction from reflective reward traces

Require: Traces $\mathcal{M}$ from a completed reference run, provider count K, temperature τ, log-multiplier cap c, minimum support $n_{\min}$

for each context and the global fallback **do**
  Select matching traces; use the global set if the context has fewer than $n_{\min}$ traces
  for each provider k **do**
    $\rho_k \leftarrow \mathrm{Corr}(s_{i,k}, y_i)$; use zero for constant scores
  end for
  $\bar{\rho} \leftarrow K^{-1} \sum_j \rho_j$
  for each provider k **do**
    $m_k \leftarrow \exp(\mathrm{clip}((\rho_k - \bar{\rho})/\tau, -c, c))$
  end for
  Store the context multipliers
end for
Export once; freeze the table for a subsequent generator-training run

evidence; conflict and subtlety gating temper disagreeing semantic assessments and intensity-sensitive cues. Their provider-specific application changes relative influence; a shared positive attenuation would largely cancel under normalization.

Composition with experience. For a fixed base weight w_k^0, define the effective contribution $a_k(x,p) = w_k^0 g_k(x,p) m_k(r(x,p))$. Reflective Reward Gating uses $m_k = 1$; EmoReward uses the fixed table produced by Evolve. With $\mathcal{A} = \{k : a_k > \epsilon_g\}$, active-provider normalization gives

$$R_{\mathrm{emo}}(x,p) = \frac{\sum_{k\in\mathcal{A}} a_k(x,p) s_k(x,p)}{\sum_{k\in\mathcal{A}} a_k(x,p) + \epsilon}. \tag{8}$$

Invalid evidence receives no mass. A common positive factor cancels except through the stabilizer; calibration acts through relative provider contributions and differences between videos. Regime r selects a table entry; gate g_k controls current evidence validity. A provider favored by memory still contributes nothing when its current evidence is invalid. With no memory entry, unit multipliers recover Reflect-only fusion; the scoring interfaces and generator objective remain the same.

3.5 Evolve: RewardMemory

RewardMemory treats reward allocation as a decision over emotional context z: the providers are actions, their fusion weights form a policy, and a stored outcome y supplies a non-parametric reliability signal. Experience is collected during a *completed reference training run* with Reflective Reward Gating. Each trace $m_i = (z_i, \{s_i^k\}, \{g_i^k\}, w_i, R_i, y_i)$ records context, scores, gates, weights, and outcomes. Evolve groups these traces into a regime neighborhood $\mathcal{N}(z)$, with a global fallback for insufficient support, and computes

$$\rho_k(z) = \frac{\mathrm{Cov}_{i\in\mathcal{N}(z)}(s_i^k, y_i)}{\sigma_{s^k}\sigma_y + \epsilon}, \qquad A_k^{\mathrm{rel}}(z) = \rho_k(z) - \frac{1}{K}\sum_j \rho_j(z). \tag{9}$$

Reliability uses raw provider scores; constant scores receive zero correlation. It measures agreement with the reference outcome, not independent emotional fidelity. The centered advantage tilts the base allocation through bounded multipliers:

$$\Delta_k = \mathrm{clip}(A_k^{\mathrm{rel}}/\tau, -c, c), \qquad w_k^{\mathrm{RM}}(z) = \frac{w_k^0 e^{\Delta_k}}{\sum_j w_j^0 e^{\Delta_j}}. \tag{10}$$

This is the closed-form optimum of a KL-regularized allocation with clipped utilities (Appendix F). The exported $m_k(r) = e^{\Delta_k}$ compose with current instance gates in Eq. 8. The table is fixed throughout the next training run: only the generator receives gradient updates, and runtime lookup neither re-estimates reliability nor writes back to the table. Algorithms 1 and 2 separate reward computation from between-run memory construction. Appendix E develops the state/action/return interpretation and context estimation; Appendix I provides a trace-based worked example.

Table 1: **EmoReward across video optimizers.** Mean ± SEM on the comparison subset defined in the experimental setting; all metrics are higher-is-better. GT is the pretrained no-RL reference. Bold marks an increase over the corresponding optimizer baseline.

Method	Overall	VQ	MQ	TA	PickScore	CLIPScore
GT (reference)	-1.1938±0.0520	-0.5899±0.0255	-0.4289±0.0220	-0.1750±0.0363	0.7059±0.0018	0.5623±0.0037
DanceGRPO	-1.2812±0.0513	-0.6392±0.0230	-0.3867±0.0235	-0.2554±0.0357	0.7177±0.0018	0.5675±0.0036
+ EmoReward	**-0.9594**±0.0574	**-0.5187**±0.0304	**-0.2054**±0.0298	**-0.2354**±0.0364	0.7039±0.0017	0.5565±0.0037
FlowGRPO	-0.7719±0.0569	-0.3567±0.0290	-0.2753±0.0271	-0.1398±0.0361	0.7061±0.0019	0.5634±0.0036
+ EmoReward	**-0.5376**±0.0556	**-0.3398**±0.0291	**-0.2302**±0.0275	**0.0323**±0.0356	**0.7115**±0.0018	0.5605±0.0037
SAGE-GRPO	-0.8269±0.0532	-0.4106±0.0261	-0.3275±0.0263	-0.0888±0.0360	0.7091±0.0019	0.5624±0.0036
+ EmoReward	**-0.2624**±0.0600	**-0.1710**±0.0304	**-0.0926**±0.0324	**0.0012**±0.0364	**0.7107**±0.0018	**0.5663**±0.0035

4 EXPERIMENTS

Experimental settings. We use HunyuanVideo 1.5 with DanceGRPO, FlowGRPO, and SAGE-GRPO, fusion $\alpha = \beta = 0.5$, batch size 2, and group size $G = 4$. The updated optimizer comparison reports average results on a subset of **302 emotional prompts** that does not overlap with the training set. This outcome-conditioned subset measures performance within the selected cases, rather than an unbiased full benchmark. GT denotes the pretrained generator without RL. Each guided row uses one fixed configuration: Reflect for DanceGRPO, and the complete EmoReward system for FlowGRPO and SAGE-GRPO; there is no per-prompt choice of the better guided output. The controlled seven-policy comparison retains its separate step-100 evaluation protocol. VideoAlign supplies R_{vid} and the reference outcome y. Memory calibration uses $K = 7$, $\tau = 0.5$, a log-multiplier cap $c = 1$, and minimum context support $n_{\min} = 20$. Appendix J.1 gives the qualitative protocol.

Evaluation metrics. Table 1 reports mean ± SEM across prompts for VideoAlign Overall, visual quality (VQ), motion quality (MQ), text alignment (TA), PickScore, and CLIPScore. The latter two are frame-averaged preference and alignment proxies; their exported logits are scaled by 26 and 30, respectively. VideoAlign is also an optimization objective, so agreement with it is distinguished from independent emotional fidelity.

4.1 EMOREWARD AS A BASE-RL-AGNOSTIC REWARD ADD-ON

Table 1 compares the pretrained reference, each optimizer baseline, and its fixed guided configuration. On the stated subset, all four VideoAlign heads increase for every optimizer. Overall improves by 0.3218 for DanceGRPO with Reflect, 0.2343 for FlowGRPO with EmoReward, and 0.5645 for SAGE-GRPO with EmoReward. The reference comparison and within-optimizer comparison answer different questions: the former evaluates the full post-training system, while the latter compares reward configurations. PickScore and CLIPScore show smaller, mixed changes; the VideoAlign gains do not imply uniform improvement on all evaluation axes.

Emotion-category analysis. Figure 4 compares GT and SAGE-GRPO with EmoReward across six metrics and eight emotion categories. The four VideoAlign heads improve in every category, while PickScore and CLIPScore show smaller, nonuniform changes. This GT comparison reflects the complete post-training system. Categories overlap within the selected subset.

4.2 ABLATIONS AND MECHANISM ANALYSIS

Agentic reward methods. Our controlled study fixes the backbone and optimization budget and varies the reward policy across seven arms: VideoAlign-only, single-provider, random, round-robin, fixed fusion, Reflective Reward Gating, and RewardMemory.

Reward-policy ablation. Table 2 compares relative reward dispersion, reliability selectivity, and normalized Emotion-Gating Benchmark (EGB) pillars. Adding RewardMemory yields the highest selectivity (0.428) and both composite scores (core 0.922, emo 0.884); the variant without memory leads the standalone preference pillar. SNR* is a dispersion statistic, not a measure of gradient quality. The composites combine an ImageReward preference proxy, dispersion, and agreement-weighted selectivity; Appendix G defines their scope.



Figure 3: **EmoReward in performance scenes.** Eight curated generated clips, each shown as three chronological full frames, illustrate expressions, gestures, and staging.

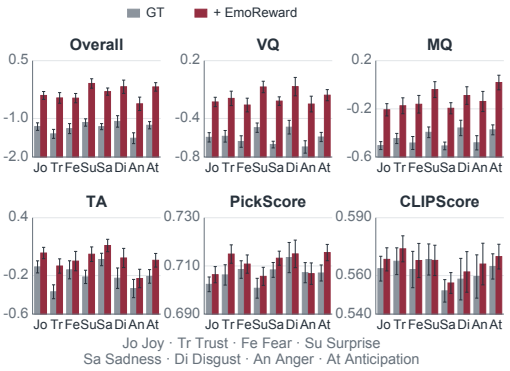


Figure 4: **Emotion-category quality.** GT (gray) versus SAGE-GRPO with EmoReward (red); mean ± SEM. Axes are truncated; categories overlap within the selected subset.



Figure 5: **Ablation of provider-pool construction.** Three frames per clip compare four configurations on the same prompt. See Appendix H for quantitative results.

Reward-provider pool. Figure 5 compares a VideoAlign baseline, semantic emotion cues, motion naturalness, and a balanced pool. Weights and curricula also differ, so this study motivates the interface rather than isolating each channel's causal effect (Appendix H). Seven channels enter memory calibration; chromatic and kinetic readouts remain auxiliary.

Memory-conditioned trust. Macro and motion receive the largest global multipliers (1.833, 1.510), whereas three constant channels receive 0.778. Memory reallocates trust but cannot activate invalid evidence. Macro shares evaluator information with the reference outcome, so its correlation reflects objective agreement. Appendix E describes context-specific allocation; Appendix I illustrates its arithmetic on real rollout scores. A matched global-memory control remains necessary to isolate the benefit of contextual calibration.

Table 2: **Reward-policy ablation under SAGE-GRPO.** Controlled policies share the backbone and training budget. Citations identify evaluators or routing/fusion mechanisms, not full-system reproductions. SNR* is relative dispersion; bold marks column maxima. Appendix G defines all metrics and the supplementary verification.

Reward method	SNR*	Sel.↑	EF↑	GD↑	GR↑	core↑	emo↑
VideoAlign (Liu et al., 2025b)	0.378	0.307	0.518	0.000	0.197	0.238	0.308
Single provider (Zhao et al., 2025)	2.403	0.314	0.454	0.305	0.147	0.302	0.340
Random routing (Wu et al., 2023a)	1.105	0.340	0.000	0.109	0.146	0.085	0.064
Round-robin (Wu et al., 2023a)	1.217	0.376	0.291	0.126	0.381	0.266	0.272
Fixed fusion (Xu et al., 2024)	0.773	0.371	0.026	0.060	0.653	0.246	0.191
EmoReward (w/o RewardMemory)	0.451	0.293	**1.000**	0.011	0.000	0.337	0.503
EmoReward (w RewardMemory)	**7.022**	**0.428**	0.767	**1.000**	**1.000**	**0.922**	**0.884**

Table 3: **Human evaluation.** Rating percentages from 18 participants on 30 prompts. Each method and criterion block contains 540 judgments. Bold marks the highest Good share per criterion; percentages are rounded to one decimal place.

	Emotion Fidelity			Temporal Coherence			Visual Quality		
Method	Bad ↓	Normal	Good ↑	Bad ↓	Normal	Good ↑	Bad ↓	Normal	Good ↑
GT	21.9%	34.8%	43.3%	23.5%	37.8%	38.7%	25.2%	38.5%	36.3%
SAGE-GRPO	31.5%	15.6%	53.0%	28.9%	19.6%	51.5%	30.9%	24.1%	45.0%
SAGE-GRPO + EmoReward	**15.4%**	24.6%	**60.0%**	**20.2%**	25.6%	**54.3%**	**15.9%**	26.9%	**57.2%**

4.3 Generalization to image backbones

Image transfer covers SD3.5 and FLUX with three optimizers. Results are objective-dependent: SD3.5–DanceGRPO improves on HPSv3 with calibration, whereas FLUX's HPSv3-optimized baseline remains strongest on that metric. Unequal SD3.5 training budgets limit causal attribution. Appendix J details the settings and full-prompt comparisons.

4.4 Human evaluation

Eighteen participants rated the three methods on 30 prompts using Bad/Normal/Good for emotion fidelity, temporal coherence, and visual quality. Table 3 shows that EmoReward obtains the highest Good shares (60.0%, 54.3%, 57.2%), compared with 53.0%, 51.5%, and 45.0% for SAGE-GRPO, and the lowest Bad shares on all three criteria. These descriptive results complement the automatic metrics; aggregation and statistical scope are detailed in Appendix B.

The complete distributions reveal a distinction that Good shares alone obscure: SAGE-GRPO also receives more Bad ratings than GT on every criterion. EmoReward lowers Bad shares relative to both comparison groups while increasing Good shares. This pattern supports examining the entire ordinal distribution, rather than reducing it to a binary preference. The reported margins do not resolve variation across people or prompts; participant- and prompt-level records are required for uncertainty estimates that account for repeated judgments.

5 Conclusion

In this paper, we proposed EmoReward, a self-evolving reward calibration framework for emotional video generation reinforcement learning. EmoReward combines instance-level Reflective Reward Gating with RewardMemory, which transfers provider-reliability estimates from a completed training run to subsequent optimization without updating reward-model parameters. Experiments across video optimizers, complemented by ablations, image-transfer analyses, and human ratings, support the utility of adaptive calibration while identifying objective-dependent trade-offs. These findings suggest that completed-run experience can inform the allocation of trust among heterogeneous reward providers, providing a foundation for further research on reward-system evolution across training.

Reproducibility Statement

The appendix specifies the reward interface, calibration rule, mathematical derivation, evaluation protocol, complete comparison tables, and qualitative examples. Source artifacts include editable figures and scripts for constructing tables and visualizations.

AI Use Statement

AI assistance was used for manuscript editing, mathematical checks, and figure preparation. The authors are responsible for verifying the scientific claims and final submission. Experimental comparison images and video frames were taken from the evaluated models; no additional media were synthesized to illustrate experimental outcomes.

References

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 59–66, 2018.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2301–2310, 2024.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. Opensearch-vl: An open recipe for frontier multimodal search agents. *arXiv preprint arXiv:2605.05185*, 2026a.

Sixiang Chen, Zhaohu Xing, Tian Ye, Xinyu Geng, Yunlong Lin, Jianyu Lai, Xuanhua He, Fuxiang Zhai, Jialin Gao, and Lei Zhu. Genevolve: Self-evolving image generation agents via tool-orchestrated visual experience distillation. *arXiv preprint arXiv:2605.21605*, 2026b.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110805–110853, 2024. URL https://arxiv.org/abs/2406.11161.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Fangming Cui, Ruixiao Zhu, Cheng Fang, Sunan Li, and Jiahong Li. A brief overview: Agentic reinforcement learning in large language models. *arXiv preprint arXiv:2604.27859*, 2026.

Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

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. *arXiv preprint arXiv:2305.16381*, 2023.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 2026.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Daya Guo et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. URL https://arxiv.org/abs/2406.15252.

Yinming Huang, Shuyuan Tu, Xi Yan, Zihan Yang, Jianhua Han, Hang Xu, Kaihang Pan, Yu-Gang Jiang, and Zuxuan Wu. Va-judger: Reward modeling from human preference feedback for joint video-audio generation. *arXiv preprint arXiv:2608.18607*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17982.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, et al. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. *arXiv preprint arXiv:2305.01569*, 2023. URL https://arxiv.org/abs/2305.01569.

Jiaqi Leng, Shuyuan Tu, Haidong Cao, Sicheng Xie, Daoguo Dong, Zuxuan Wu, and Yu-Gang Jiang. Preference score distillation: Leveraging 2d rewards to align text-to-3d generation with human preference. *arXiv preprint arXiv:2603.01594*, 2026.

Yibo Li, Zijie Lin, Ailin Deng, Xuan Zhang, Yufei He, Shuo Ji, Tri Cao, and Bryan Hooi. Just-in-time reinforcement learning: Continual learning in llm agents without gradient updates. *arXiv preprint arXiv:2601.18510*, 2026.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. *arXiv preprint arXiv:2501.16566*, 2025. URL https://arxiv.org/abs/2501.16566.

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. URL https://arxiv.org/abs/2501.13918.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. VideoDPO: Omni-Preference Alignment for Video Diffusion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c. URL https://arxiv.org/abs/2412.14167.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2310.11440.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-Level Reward Design via Coding Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.12931.

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. URL https://arxiv.org/abs/2508.03789.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://arxiv.org/abs/2303.17651.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980.

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. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://arxiv.org/abs/2303.11366.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2407.14505.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. StableAvatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025a.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. StableAnimator: High-quality identity-preserving human image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2025b.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

Huanzhen Wang, Ziheng Zhou, Jiaqi Song, Li He, Yunshi Lan, Yan Wang, and Wenqiang Zhang. ARGen: Affect-Reinforced Generative Augmentation towards Vision-based Dynamic Emotion Perception, 2026a. URL https://arxiv.org/abs/2604.12255.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.

Yibin Wang, Yuhang Zang, Feng Han, Jiazi Bu, Yujie Zhou, Cheng Jin, and Jiaqi Wang. Unified Personalized Reward Model for Vision Generation. *arXiv preprint arXiv:2602.02380*, 2026b. URL https://arxiv.org/abs/2602.02380.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023a. URL https://arxiv.org/abs/2308.08155.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis, 2023b. URL https://arxiv.org/abs/2306.09341.

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. *arXiv preprint arXiv:2304.05977*, 2023. URL https://arxiv.org/abs/2304.05977.

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. URL https://arxiv.org/abs/2412.21059.

Ran Xu, Tianci Liu, et al. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training. *arXiv preprint arXiv:2602.01511*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic Memory for LLM Agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.12110.

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.

Yu Yang, Yue Liao, Jianbiao Mei, Baisen Wang, Xuemeng Yang, Licheng Wen, Jiangning Zhang, Xiangtai Li, Liang Lv, Hanlin Chen, Botian Shi, Yong Liu, Shuicheng Yan, and Gim Hee Lee. SPIRAL: Self-Evolving Action-Conditioned Video Generation via Reflective Planning Agents, 2026. URL https://arxiv.org/abs/2603.08403.

Qiying Yu et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-Rewarding Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2401.10020.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025.

Rongzhi Zhang, Rui Feng, Zhihan Zhang, Jingfeng Yang, Qingyu Yin, Xin Liu, Zixuan Zhang, Priyanka Nigam, Bing Yin, Tuo Zhao, and Chao Zhang. Qubric: Co-designing queries and rubrics for rl beyond verifiable rewards. *arXiv preprint arXiv:2606.03968*, 2026.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-Omni: Explainable Omni-Multimodal Emotion Recognition with Reinforcement Learning. *arXiv preprint arXiv:2503.05379*, 2025. URL https://arxiv.org/abs/2503.05379.

Mingzhe Zheng, Weijie Kong, Yue Wu, Dengyang Jiang, Yue Ma, Xuanhua He, Bin Lin, Kaixiong Gong, Zhao Zhong, Liefeng Bo, et al. Manifold-aware exploration for reinforcement learning in video generation. *arXiv preprint arXiv:2603.21872*, 2026.

A Limitations and Future Work

Beyond emotional video generation. EmoReward studies a training-adaptive reward system that transfers evidence about provider reliability between generator-training runs without updating reward-model parameters. The present evaluation focuses on emotion, where different scenes call for different combinations of facial, gestural, semantic, and motion evidence. A broader application would require a task-appropriate provider pool and a meaningful outcome for calibration. Action fidelity and audio–visual correspondence are natural visual extensions; language models and tool-use agents could similarly combine semantic judges, execution checks, and trajectory-level assessments. Whether the same calibration rule transfers effectively remains an empirical question, especially when providers share errors or outcomes favor only part of the task.

Iterative evolution across training runs. A completed training run can be treated as a unit of experience for its successor. Repeating this procedure could produce a sequence of generators and reward-allocation tables, with each generation contributing new traces. Our experiments evaluate a reference-to-calibrated transition, not sustained improvement across many generations. Such an extension must address distribution shift, stale evidence, and feedback loops that reinforce a proxy's errors. Retaining diverse historical traces and evaluating held-out outcomes would be important tests of whether repeated calibration remains useful.

Model-level human feedback. The trace representation also suggests a way to carry human assessments of a completed model into subsequent training. For example, a reviewer who finds movement too weak could request greater emphasis on suitable motion providers in the affected contexts; excessive motion could motivate the opposite adjustment. This would complement sample-level preference feedback with structured assessments of a model's final behavior, translated into bounded changes to the next run's reward allocation. Mapping subjective feedback to providers and regimes, reconciling different raters, and preventing compensating failures remain open design questions. The ordinal study reported here evaluates outputs; it does not implement this proposed feedback mechanism.

Evaluation scope. The optimizer comparison uses an outcome-conditioned subset, provider construction changes both weights and curricula, and some image comparisons have unequal training budgets. VideoAlign agreement is not an independent measure of emotional fidelity. Human ratings provide complementary descriptive evidence, but aggregate counts cannot support participant- or prompt-aware uncertainty estimates. These limits motivate unselected evaluations, matched-budget controls, and fully documented human studies.

B Human Evaluation: Ordinal Ratings

Study and comparison groups. The study includes 18 participants and 30 prompts, with one video per prompt from each of three methods: the pretrained generator without RL (GT), SAGE-GRPO, and SAGE-GRPO with EmoReward. GT is a generator baseline, not a ground-truth human video. Participants assessed each output separately on three criteria using Bad, Normal, or Good. The supplied aggregate results contain 540 judgments for each method–criterion combination, consistent with the full participant–prompt product. Across all three methods and criteria, there are 4,860 categorical judgments. These ordinal counts are separate from the earlier pairwise preference protocol; pairwise choices are not converted into categorical ratings.

Criteria. *Emotion fidelity* concerns how convincingly the video conveys the prompt's intended affect, including whether expression and gesture fit its emotional tone. *Temporal coherence* concerns continuity of expression, identity, and movement through the clip. *Visual quality* concerns perceptual detail, anatomy, and visible artifacts. The categories distinguish poor, intermediate, and good quality on each axis. They are rated separately rather than collapsed into a single preference or an assumption that stronger emotional intensity is always better.

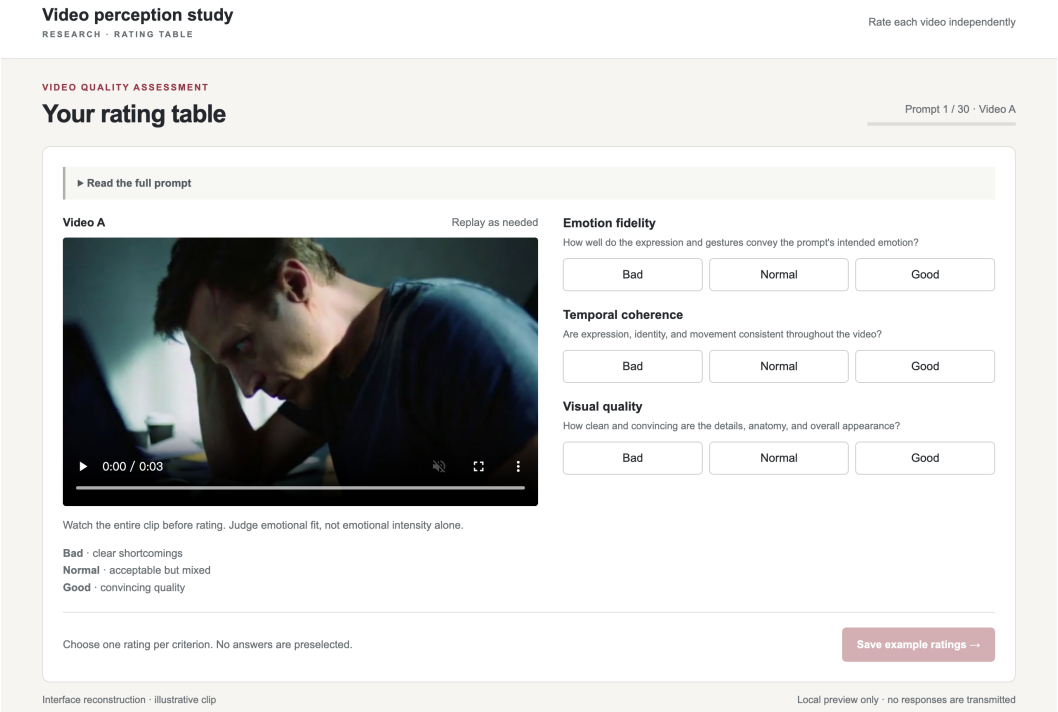


Figure 6: **Interface for ordinal human evaluation.** The interface presents the prompt and generated video with independent ratings of emotional fidelity, temporal coherence, and visual quality, each using three ordered categories: Bad, Normal, and Good. This reconstruction illustrates the reported rating protocol rather than the original data-collection interface.

Aggregation and rounding. For method m, criterion c, and rating category b, the percentage is:

$$\widehat{p}_{m,c,b} = 100\,\frac{N_{m,c,b}}{\sum_{b' \in \{\mathrm{Bad},\mathrm{Normal},\mathrm{Good}\}} N_{m,c,b'}}, \qquad \sum_{b'} N_{m,c,b'} = 540. \tag{11}$$

Table 3 rounds each percentage independently to one decimal place and bolds only the highest Good percentage within each criterion. Rounding may make a three-category total differ from 100.0% by 0.1 percentage points. The raw integer counts are preserved in the accompanying source data, rather than rounded before normalization. All nine aggregate blocks have the stated denominator; the totals alone do not establish participant-level completeness or the absence of duplicate records.

Reading the distributions. Relative to SAGE-GRPO, EmoReward's Good share increases by 7.0, 2.8, and 12.2 percentage points for emotion fidelity, temporal coherence, and visual quality, respectively. These differences are computed from the unrounded counts. Its Bad share falls from 31.5% to 15.4% for emotion fidelity, from 28.9% to 20.2% for temporal coherence, and from 30.9% to 15.9% for visual quality. The complete distribution matters: compared with GT, SAGE-GRPO has more Good ratings but also more Bad ratings on every criterion, so a Good-only summary would omit part of the observed pattern.

Statistical scope. The present report uses the supplied aggregated counts. It does not treat repeated judgments from the same participant or prompt as independent observations. Participant- and prompt-level records are needed for uncertainty estimates that respect both sources of dependence; no confidence interval or significance claim is inferred from the aggregate margins. The aggregates do not document recruitment, presentation order, blinding, exclusions, or prompt-selection provenance. Those collection details must be established from the study records rather than inferred from the rating totals.

Table 4: **Primitive family, provider, backing estimator, and defining equation.** The two appearance channels (chromatic, kinetic) are always-on lightweight signals; the remaining seven form the memory support whose per-provider reliability is reported in the calibration analysis (Table 6).

Family	Provider	Estimator / signal	Eq.
Facial	AU matching	Facial action-unit head	12
Facial	CLIP-AUTT	Temporal AU–text similarity	13
Semantic	R1-Omni	Valence–arousal estimation	14
Semantic	Macro	Alignment and AU compliance	15
Body	Body cues	Body-marker satisfaction	16
Temporal	Peak alignment	AU peak versus target window	18
Appearance	Chromatic	Saturation and warm hue	19
Appearance	Kinetic	Inter-frame differences	20
Motion	Naturalness	Optical-flow consistency	21

C ACT: PROVIDER INTERFACE AND REFERENCE SCORING RULES

This appendix specifies the scoring operators used by the provider interface in Eq. 6. The six primitive *families* (facial expression; semantic emotion understanding; body and temporal evidence; chromatic atmosphere; kinetic intensity; motion naturalness) are instantiated by nine observation channels; Table 4 maps each family to its provider(s), the backing estimator, and the defining equation. Throughout, T is the number of sampled frames, $\mathcal{F} \subseteq \{1, \dots, T\}$ the face-detected subset, ω_t the EmoScript time weight (per-window kernels $\psi_\ell(t) = \mathbf{1}[t \in \omega_\ell]/(\sum_{t'} \mathbf{1}[t' \in \omega_\ell] + \epsilon)$ over $\Omega = \{[t_s^\ell, t_e^\ell]\}$), and $\kappa(x; \sigma) = \exp(-\max(0, x)/\sigma)$ a bounded exponential score. Each provider k returns a score $s_k \in [0, 1]$ and a validity signal u_k.

C.1 FACIAL EXPRESSION: AU MATCHING AND EXPRESSION SIMILARITY

AU matching (*micro*). A face detector followed by a multi-task action-unit head emits eight AU intensities per detected face, $\widehat{q}_{j,t} \in [0, 1]$ for $j \in \mathcal{A}_{\mathrm{FACS}} = \{1, 2, 4, 6, 9, 12, 25, 26\}$. Against the EmoScript target intensities $q_j^\star$ we score agreement by a per-AU intersection-over-union overlap (rather than an ℓ_p kernel, so a target and an observation are rewarded for occupying the same activation interval), averaged over face frames:

$$\mathrm{IoU}(q_j^\star, \widehat{q}_{j,t}) = \frac{\min(q_j^\star, \widehat{q}_{j,t})}{\max(q_j^\star, \widehat{q}_{j,t}) + \epsilon}, \quad s_{\mathrm{micro}} = \frac{1}{|\mathcal{A}^\star|} \sum_{j \in \mathcal{A}^\star} \frac{1}{|\mathcal{F}|} \sum_{t \in \mathcal{F}} \mathrm{IoU}(q_j^\star, \widehat{q}_{j,t}), \tag{12}$$

with AU compliance and validity $c_{\mathrm{AU}} = \frac{1}{|\mathcal{A}^\star_{\mathrm{req}}|} \sum_{j \in \mathcal{A}^\star_{\mathrm{req}}} \mathbf{1}[\max_{t \in \mathcal{F}} \widehat{q}_{j,t} > \theta_{\mathrm{AU}}]$ and $u_{\mathrm{micro}} = |\mathcal{F}|/T$. The provider is inactive when no frame yields a face with AU readings. In our implementation, $\theta_{\mathrm{AU}} = 0.3$.

Expression similarity (*clip_autt*). A temporal vision–language encoder embeds a T_{clip}-frame clip into $\mathbf{v}$ and compares it to N_{au} action-unit text-prompt embeddings $\{\mathbf{t}_j\}$ by cosine similarity, mapping the mean similarity from $[-1, 1]$ to $[0, 1]$:

$$\bar{s}_{\mathrm{au}} = \frac{1}{N_{\mathrm{au}}} \sum_j \cos(\mathbf{v}, \mathbf{t}_j), \qquad s_{\mathrm{clip}} = \tfrac{1}{2}(\bar{s}_{\mathrm{au}} + 1). \tag{13}$$

An auxiliary discrete head additionally yields a negative-affect probability $p_{\mathrm{neg}} = \mathrm{softmax}(\mathrm{logits})_{\mathrm{neg}}$, from which the scorer derives $\hat{v} = 1 - p_{\mathrm{neg}}$ and $\hat{a} = \min(1, \lambda_a p_{\mathrm{neg}} + b_a)$; the linear rescale in Eq. 13 (rather than clip) preserves the ordering on clips where $\bar{s}_{\mathrm{au}} < 0$. In our implementation, $T_{\mathrm{clip}} = 16$, $N_{\mathrm{au}} = 47$, $\lambda_a = 1.5$, and $b_a = 0.2$.

C.2 SEMANTIC EMOTION UNDERSTANDING: VALENCE–AROUSAL READING AND THE ALIGNMENT SEED

Valence–arousal reading (*r1omni*). A multimodal emotion-understanding model produces a free-text emotion description that is parsed to an emotion label and continuous valence–arousal estimates

$(\hat{v}, \hat{a})$. Given EmoScript targets $(v^\star, a^\star)$ the reward supports three modes (default: valence):

$$s_{\mathrm{r1}} = \begin{cases} \frac{1}{2}\big[(1-|\hat{v}-v^\star|)+(1-|\hat{a}-a^\star|)\big], & \textsc{alignment}, \\ \hat{v}, & \textsc{valence}, \\ \frac{1}{2}(\hat{v}+\hat{a}), & \textsc{mean}. \end{cases} \tag{14}$$

Text-alignment seed (*macro*). When a general video–text alignment model is available it blends its text-alignment head $\mathrm{TA} \in [0,1]$ with AU compliance weighted by face rate $w_{\mathrm{face}} = \min(\text{face rate}, 1)$; otherwise a heuristic reads compliance directly:

$$s_{\mathrm{macro}} = \begin{cases} \lambda_{\mathrm{TA}}\,\mathrm{TA} + (1-\lambda_{\mathrm{TA}})c_{\mathrm{AU}}w_{\mathrm{face}}, & \text{VideoAlign reflector}, \\ b_{\mathrm{macro}} + (1-b_{\mathrm{macro}})c_{\mathrm{AU}}\pi_{\mathrm{conf}}, & \text{heuristic (active iff face rate} \geq \theta_{\mathrm{macro}}). \end{cases} \tag{15}$$

In our implementation, $\lambda_{\mathrm{TA}} = 0.6$, $b_{\mathrm{macro}} = 0.3$, and $\theta_{\mathrm{macro}} = 0.1$. The confidence π_{conf} belongs to the heuristic reflector. Because the alignment-seeded branch shares an evaluator with the reference outcome, its correlation cannot serve as independent emotion validation.

C.3 Body and temporal evidence (body, temporal)

The body provider counts each EmoScript body marker $m \in \mathcal{M}^\star$ satisfied by at least one frame; the temporal provider locates the frame $t^\star$ of peak expected-AU energy and scores its distance to the EmoScript window $[t_s, t_e]$:

$$s_{\mathrm{body}} = \frac{1}{|\mathcal{M}^\star|}\sum_{m\in\mathcal{M}^\star} \mathbf{1}\big[\exists\, t:\ b_{m,t}=1\big], \tag{16}$$

$$t^\star = \arg\max_t \sum_{j\in\mathcal{A}^\star} \widehat{q}_{j,t}, \tag{17}$$

$$s_{\mathrm{temporal}} = \begin{cases} 1, & t_s \leq t^\star \leq t_e, \\ \max\!\big(0,\ 1 - \dfrac{\min(|t^\star - t_s|, |t^\star - t_e|)}{\max(t_e - t_s,\ \delta_\Omega)}\big), & \text{else}. \end{cases} \tag{18}$$

The window stabilizer is $\delta_\Omega = 0.1$ in our implementation. These scores require their corresponding observations; unavailable evidence receives the neutral fallback s_{neutral} and validity zero. We use $s_{\mathrm{neutral}} = 0.5$. In the analyzed traces, the micro, body, and temporal channels are constant at 0.5; this does not imply that every generated video lacks a face or body.

C.4 Chromatic atmosphere (chromatic)

A per-frame chromatic signal blends the mean saturation $\bar{S}_t \in [0,1]$ with the fraction of warm-hue pixels $W_t = \overline{\mathbf{1}[h \leq \theta^-_{\mathrm{warm}} \vee h \geq \theta^+_{\mathrm{warm}}]}$ (on the standard $[0, 179]$ hue scale), and the clip-level score aligns the frame mean to the target valence $v^\star$ (linearly mapped to $[0,1]$):

$$c_t = \lambda_{\mathrm{sat}}\bar{S}_t + (1-\lambda_{\mathrm{sat}})W_t, \qquad \bar{c} = \tfrac{1}{T}\textstyle\sum_t c_t, \qquad s_{\mathrm{chromatic}} = \mathrm{clip}_{[0,1]}(1 - |\bar{c} - v^\star|). \tag{19}$$

In our implementation, $\lambda_{\mathrm{sat}} = 0.5$ and the warm-hue thresholds are $\theta^-_{\mathrm{warm}} = 30$, $\theta^+_{\mathrm{warm}} = 150$ on the stated hue scale.

C.5 Kinetic intensity (kinetic)

A per-frame kinetic signal normalizes the mean absolute frame difference using saturation scale σ_{kin}, and the clip-level score aligns the frame mean to the target arousal $a^\star$:

$$k_t = \mathrm{clip}_{[0,1]}\big(\tfrac{1}{\sigma_{\mathrm{kin}} B_I}\,\overline{|I_t - I_{t-1}|}\big), \qquad \bar{k} = \tfrac{1}{T}\textstyle\sum_t k_t, \qquad s_{\mathrm{kinetic}} = \mathrm{clip}_{[0,1]}(1 - |\bar{k} - a^\star|). \tag{20}$$

In our implementation, $\sigma_{\mathrm{kin}} = 0.20$ and the pixel range is $B_I = 255$. Appearance features describe potential cues; they are not part of the seven-provider memory update.

C.6 MOTION NATURALNESS (MOTION)

Dense optical flow $\mathbf{f}_t$ on grayscale, short-side-L_{flow} frames yields per-frame magnitude $m_t = \overline{\|\mathbf{f}_t\|}$ and a global (dominant) motion vector $\mathbf{g}_t = \overline{\mathbf{f}_t}$. Four bounded sub-scores reward smooth flow, low camera jerk, low warp residual, and low compensated flicker:

$$\begin{aligned} r_{\text{flow}} &= \kappa(\overline{|\Delta m_t|};\, \sigma_{\text{flow}}), & r_{\text{cam}} &= \kappa(\overline{\|\Delta^2 \mathbf{g}_t\|};\, \sigma_{\text{cam}}), \\ r_{\text{warp}} &= \kappa(\overline{|\,\mathcal{W}(I_{t-1}; \mathbf{f}_t) - I_t\,|};\, \sigma_{\text{warp}}), & r_{\text{flick}} &= \kappa(\overline{|\,\mathcal{S}(I_{t-1}; \mathbf{g}_t) - I_t\,|};\, \sigma_{\text{flick}}), \end{aligned} \tag{21}$$

where $\mathcal{W}$ is flow-warping and $\mathcal{S}$ a global-shift compensation. With base $r_{\text{base}} = \sum_{j \in \{\text{flow,cam,warp,flick}\}} \lambda_j r_j$ and activity $E = 1 - \exp(-\bar{m}/\sigma_E)$, the arousal-conditioned score and confidence are

$$s_{\text{motion}} = \lambda_M r_{\text{base}} + (1 - \lambda_M)(1 - |E - a^\star|), \tag{22}$$

$$u_{\text{motion}} = \lambda_U \min\!\left(1, \tfrac{T-2}{n_U}\right) + (1 - \lambda_U) \min\!\left(1, \tfrac{\text{std}(m_t)}{\sigma_U}\right), \tag{23}$$

reverting to $s_{\text{motion}} = r_{\text{base}}$ when no target arousal is supplied. The motion gate scales linearly with this observation confidence (Appendix D). The mixture coefficients obey $\sum_j \lambda_j = 1$. In our implementation, $L_{\text{flow}} = 160$, $(\sigma_{\text{flow}}, \sigma_{\text{cam}}, \sigma_{\text{warp}}, \sigma_{\text{flick}}) = (0.35, 0.65, 0.085, 0.11)$, and $(\lambda_{\text{flow}}, \lambda_{\text{cam}}, \lambda_{\text{warp}}, \lambda_{\text{flick}}) = (0.34, 0.24, 0.24, 0.18)$. We use $\sigma_E = 1.5$, $\lambda_M = 0.82$, $\lambda_U = 0.6$, $n_U = 8$, and $\sigma_U = 0.035$.

C.7 SCORE NORMALIZATION AND VALIDITY

The evaluated fusion consumes the providers' bounded scores directly. An optional running-statistic normalization is not an additional transformation applied to those scores. Availability and confidence are retained separately from score magnitude: unavailable evidence receives a neutral fallback but zero effective contribution. Motion confidence is incorporated once through its channel gate, while an activity mask handles unavailable motion observations. This convention avoids counting the same confidence twice. The micro, body, and temporal channels are constant in the analyzed reference traces; their neutral readings do not establish the absence of visible faces or bodies in the generated videos.

D REFLECT: INSTANCE-DEPENDENT GATE CONSTRUCTION

Reflect separates training progress, semantic consistency, observation availability, and the prompt's intended intensity. A factor is neutral for providers to which it does not apply. Let t denote the generator update, $s_0 < s_1$ the phase-transition endpoints, $\gamma_{\min}$ a phase floor, and q_f the available face-evidence ratio:

$$h(t) = \text{clip}_{[0,1]}\left(\frac{t - s_0}{s_1 - s_0}\right), \quad b = \max(\gamma_{\min}, h(t)) g_{\text{conflict}}, \quad f = \min(1, q_f/\theta_f). \tag{24}$$

The semantic-conflict factor attenuates disagreement $d = |s_{\text{CLIP-AUTT}} - s_{\text{R1-Omni}}|$ above a threshold θ_d, with lower bound γ_d. Subtlety is detected from prompt cues or $a^\star \le \theta_a$; it selects restraint factors ξ_f, ξ_t, and ξ_e for facial/body, temporal, and semantic channels, respectively. Otherwise these restraint factors equal one. Before applying observation-availability masks, the gates are

$$\begin{aligned} g_{\text{macro}} &= b, & g_{\text{micro}} &= g_{\text{body}} = bf\xi_f, & g_{\text{temporal}} &= bf\xi_t, \\ g_{\text{CLIP-AUTT}} &= g_{\text{R1-Omni}} = b\xi_e, & g_{\text{motion}} &= \lambda_g + (1 - \lambda_g) c_{\text{motion}}. \end{aligned} \tag{25}$$

Here c_{motion} is the optical-flow observation confidence. Availability masking and continuous confidence together instantiate the validity term in Eq. 7; confidence is not multiplied a second time.

Implementation settings. In our implementation, the default warmup endpoints are $s_0 = 40$ and $s_1 = 100$, with $\gamma_{\min} = 0.1$. Face gating uses $\theta_f = 0.35$; semantic conflict uses $\theta_d = 0.20$ and $\gamma_d = 0.35$. For subtle prompts, detected using $\theta_a = 0.4$ or prompt cues, $(\xi_f, \xi_t, \xi_e) = (0.85, 0.75, 0.80)$; the motion floor is $\lambda_g = 0.55$. These are rule-based controls rather than learned reward-model parameters. Historical exported traces retain their recorded gates; the default specification alone does not reconstruct a past run's full configuration.

Effect on normalized fusion. A factor shared by every active provider cannot by itself redistribute trust. For $g_k = \gamma h_k$ with $\gamma > 0$, active normalization yields

$$R_{\text{emo}} = \frac{\sum_k w_k h_k s_k}{\sum_k w_k h_k + \epsilon/\gamma}. \tag{26}$$

Except for the stabilizer, the common scale cancels. Reflect changes the relative allocation through provider-specific factors, invalidation, and composition with the original reward branch. The face-activity proxy used to group exported traces in Appendix E is distinct from the current video's face-evidence ratio.

E EVOLVE: EXPERIENCE, STATE, AND CONTEXT ESTIMATION

Reward allocation as an agentic decision. The *state* is the context signature z of a prompt–video pair: target affect, face-branch activity, motion, subtlety, and provider validity. The *action set* consists of the reward primitives; allocating mass among them determines which observations influence training. The *return* y is the recorded evaluation outcome associated with a generated video. The *policy* $z \mapsto w(z)$ is thus a context-dependent distribution of reward trust, not the video generator itself. RewardMemory acts as a *non-parametric critic*: it retains reference-run traces and estimates how each provider tracks outcomes in similar contexts without fitting a value network. The correspondence to $A = Q - V$ centers provider reliability against the across-provider baseline. It is an allocation interpretation, not a claim that Pearson correlation is a Bellman-consistent value estimate.

Trace construction and context support. Experience is collected during a completed reflective reference run. Each evaluated video contributes $m_i = (z_i, \{s_i^k\}, \{g_i^k\}, w_i, R_i, y_i)$. Scores, gates, and applied weights preserve different aspects of the reward call; reliability estimation uses the raw scores and outcome, not gate-multiplied scores.

The state signature contains target valence and arousal, face-branch activity, motion, and subtlety. The context construction maps positive/negative valence keywords to $0.8/0.2$, with neutral fallback 0.5, and uses analogous high/low arousal targets. Face-branch activity is one when a selected facial, body, temporal, or AU-related branch moves away from its neutral default, and zero otherwise. It is not a face-pixel fraction or the fraction of frames containing a face.

Memory construction partitions traces with thresholds $\theta_{\text{reg},f}$ for face activity and $\theta_{\text{reg},a}$ for target arousal. In our implementation, $\theta_{\text{reg},f} = 0.35$ and $\theta_{\text{reg},a} = 0.5$. The analyzed sample populates two face-activity-zero regimes, with 2,724 lower-arousal and 508 higher-arousal traces. A regime uses its own estimate when it contains at least $n_{\min}$ traces, otherwise falling back to the global calibration. We use $n_{\min} = 20$. The exported table uses these regime subsets. The memory helper also supports offline similarity-based neighborhoods $\mathcal{N}_L(z)$; L denotes neighborhood size and is distinct from the provider count $K = 7$.

Between the lower- and higher-arousal regimes, the macro multiplier increases from 1.807 to 1.990, while motion decreases from 1.567 to 1.204. These are observed weight-allocation changes; their downstream benefit requires a matched global-memory control.

Reliability is the Pearson correlation between the raw provider score and VideoAlign overall within the selected context. Constant scores receive correlation zero by convention. The centered reliability advantage determines an unnormalized multiplier, which composes with the base weight and Reflect factor. The table is constructed after the reference training run and held fixed while the subsequent calibrated generator is trained. Runtime evaluation performs a regime lookup, not nearest-neighbor retrieval. Online refresh is an extension of this procedure, rather than a property established by a fixed-table experiment.

The two calibration variants. This design yields the two calibration variants evaluated in Section 4. *Reflective Reward Gating* uses the Reflect stage alone, producing the instance-calibrated base weights with $m_k = 1$; *RewardMemory* composes that base with the Evolve tilt of Eq. 10. This composition—not the reflective gate in isolation—is the mechanism we evaluate. The reference-run memory construction in Algorithm 2 details the Evolve stage.

F Derivation of the Closed-Form KL Update

For positive base weights on the supported simplex, consider

$$\max_{w \in \Delta^{K-1}} \sum_k w_k A_k^{\mathrm{rel}} - \tau \sum_k w_k \log \frac{w_k}{w_k^0}. \tag{27}$$

The Lagrangian has stationary condition $A_k^{\mathrm{rel}} - \tau(\log(w_k/w_k^0)+1)+\lambda = 0$. Normalizing the resulting exponentials gives $w_k = w_k^0 e^{A_k^{\mathrm{rel}}/\tau} / \sum_j w_j^0 e^{A_j^{\mathrm{rel}}/\tau}$. Strict concavity yields a unique optimum. A shared additive baseline in the utilities cancels from this unclipped solution.

For $\Delta_k = \mathrm{clip}(A_k^{\mathrm{rel}}/\tau, -c, c)$, the same derivation applies to clipped utilities $\widetilde{A}_k = \tau\Delta_k$. Since $e^{-c} \le \sum_j w_j^0 e^{\Delta_j} \le e^c$, normalized weights satisfy

$$e^{-2c} \le \frac{w_k^{\mathrm{RM}}}{w_k^0} \le e^{2c}. \tag{28}$$

The cap bounds relative changes; it does not guarantee a one-hot distribution as $\tau \to 0$. Zero base weights remain outside the support.

Scope of the bound and cold start. The interval $[e^{-c}, e^c]$ bounds the exported, unnormalized multipliers; the normalized weight ratio has the wider bound above. Clipping the utilities is not equivalent to solving the original objective subject to a prescribed hard KL constraint. If all advantages are zero, the update returns w^0. A constant provider has $\rho_k = 0$, but its centered advantage is negative only when the mean reliability is positive; constant scores therefore do not imply unconditional down-weighting. In the supplied traces, positive-reliability providers make this relative attenuation observable. None of these algebraic properties proves improved generalization or emotional fidelity.

Group-relative invariance. If $R_i' = aR_i + b$ with $a > 0$, then $\mu' = a\mu + b$, $\sigma' = a\sigma$, and

$$A_i' = \frac{R_i - \mu}{\sigma + \epsilon/a}. \tag{29}$$

Translation cancels exactly. By contrast, $\mathrm{CV}(R) = \sigma/(|\mu| + \epsilon)$ is not translation invariant. Consequently, a reward offset can increase relative dispersion without improving the standardized group advantage. This is an algebraic property, not an experimental claim about learning efficiency.

Generator Optimization Objective For completeness, we state the clipped GRPO objective underlying the group advantages in Eq. 1 (Shao et al., 2024). Let $h_{i,t}$ denote the generation state and $a_{i,t}$ its sampled transition in candidate trajectory i. Define $\varrho_{i,t}(\theta) = \pi_\theta(a_{i,t} \mid h_{i,t}, p)/\pi_{\theta_{\mathrm{old}}}(a_{i,t} \mid h_{i,t}, p)$. With clipping radius $\varepsilon_{\mathrm{clip}}$ and reference-policy regularization coefficient κ,

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{p,\{x_i\}\sim\pi_{\theta_{\mathrm{old}}}} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} \big(\min\{\varrho_{i,t}A_i, \mathrm{clip}(\varrho_{i,t}, 1-\varepsilon_{\mathrm{clip}}, 1+\varepsilon_{\mathrm{clip}})A_i\} - \kappa\, D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h_{i,t}, p) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid h_{i,t}, p))\big)\Bigg]. \tag{30}$$

Here T_i is the number of sampled transitions, $\pi_{\theta_{\mathrm{old}}}$ is the rollout policy, and π_{ref} is a fixed reference policy. The scalar reward of the completed video determines A_i; advantages and reward-provider outputs are treated as constants during the policy update. This expression states the common clipped objective; individual video optimizers specify their transition likelihoods, sampled timesteps, and regularization. EmoReward changes the scalar reward supplied to the optimizer, not those optimizer-specific choices. Its generator KL penalty is distinct from the reward-allocation KL term derived above, whose variables are provider weights rather than generator parameters.

G Reward-Gating Metrics and Reproducibility

Table 2 reports two reward diagnostics (SNR* and Sel.), three normalized pillars (EF, GD, GR), and two composites (core, emo). All seven policies use the same backbone and optimization budget. The exported evaluation contains 160 four-candidate groups for VideoAlign, single-provider, and random routing, and 152 groups for each remaining policy. Thus the available sample counts differ, even though the nominal training budget is shared. The references identify scoring models or routing mechanisms, not reproductions of complete agentic systems.

G.1 Reward Diagnostics

Let R_i denote the fused training reward, y_i the VideoAlign reference outcome, and $s_{i,k}$ provider k's score. A group g contains candidates for one prompt.

Reward–outcome agreement (ROA). ROA is an auxiliary input to GR rather than a displayed column in Table 2. It measures within-group pairwise ranking agreement. For outcome-nontied pairs $\mathcal{C}_g = \{(i,j) : i < j,\ |y_i - y_j| > \epsilon_y\}$, the intended statistic is

$$\mathrm{Agr}(g) = \frac{1}{|\mathcal{C}_g|} \sum_{(i,j)\in\mathcal{C}_g} \left[\mathbf{1}\{(R_i - R_j)(y_i - y_j) > 0\} + \frac{1}{2}\mathbf{1}\{R_i = R_j\} \right]. \tag{31}$$

The reported agreement is $\mathrm{ROA} = \mathrm{Mean}_g\, \mathrm{Agr}(g)$. The supplied pairwise helper uses $\epsilon_y = 10^{-9}$ and returns $1/2$ when no outcome-nontied pairs exist. However, its treatment of reward ties differs from the symmetric half-credit convention above. The original aggregation script is unavailable; we therefore retain the archived ROA values as inputs to GR, without claiming to have recomputed their tie handling. VideoAlign-only obtains one by construction when its reward and reference ordering coincide. ROA measures agreement with that evaluator, not independent human or emotional quality.

Within-group relative dispersion (SNR*). With population standard deviation within each group,

$$D_g = \frac{\mathrm{Std}_{i\in g}(R_i)}{|\,\mathrm{Mean}_{i\in g}(R_i)| + \epsilon_D}, \qquad \mathrm{SNR}^* = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} D_g. \tag{32}$$

The supplied diagnostic implementation uses $\epsilon_D = 10^{-6}$. This is a coefficient of variation, not a signal-to-noise power ratio. A reward offset can increase it while leaving centered GRPO advantages unchanged; a high value neither establishes nor is necessary for better policy gradients. We report the archived per-policy reductions, separately from the aggregate-level checks below.

Reliability selectivity (Sel.). Let $\rho_k = \mathrm{Corr}(s_{\cdot,k}, y)$ over the policy's evaluated candidates, using the nine-channel diagnostic pool $\mathcal{P}$. For $Z = \sum_{k\in\mathcal{P}} |\rho_k| > 0$,

$$q_k = \frac{|\rho_k|}{Z}, \qquad \mathrm{Sel.} = 1 - \frac{-\sum_{k\in\mathcal{P}} q_k \log q_k}{\log |\mathcal{P}|}, \qquad 0\log 0 = 0. \tag{33}$$

The pool includes chromatic and kinetic readouts as well as the seven fusion providers. Sel. describes concentration of absolute provider–outcome correlations, not applied fusion weights. It discards correlation signs and does not establish causal provider usefulness. All seven reported policies have $Z > 0$; the auxiliary zero-sample row and its all-zero-profile convention are excluded. Recomputing Sel. from the exported correlation profiles agrees with the stored values within 5×10^{-5}.

G.2 Normalized Pillars and Composite Scores

For the fixed comparison set $\mathcal{A}$ of seven policies, define

$$N(v_a) = \frac{v_a - \min_{b\in\mathcal{A}} v_b}{\max_{b\in\mathcal{A}} v_b - \min_{b\in\mathcal{A}} v_b}. \tag{34}$$

The supplied reducer returns zero for a constant input across policies; none of the three reported inputs is constant. Normalization is specific to the compared set, so the resulting scores are not absolute measures transferable across benchmarks.

EF: preference-proxy pillar. $\mathrm{EF}_a = N(I_a)$, where I_a is the archived mean ImageReward score (Xu et al., 2023). EF retains the benchmark's "Emotion Fidelity" abbreviation, but its input is a general visual preference model. We do not interpret it as an independent measurement of emotion fidelity. The raw means and normalization inputs are included in the supplement.

GD: relative-dispersion pillar. $\mathrm{GD}_a = N(\mathrm{SNR}^*_a)$. The benchmark calls this Gating Discriminativeness; mathematically, it is the normalized coefficient of variation above. Its inclusion does not turn reward dispersion into a gradient-quality measurement.

GR: agreement-weighted selectivity pillar. $\mathrm{GR}_a = N(\mathrm{Sel.}_a\,\mathrm{ROA}_a)$. GR combines correlation-profile concentration with agreement with the reference evaluator. Both inputs are diagnostic summaries; the product does not establish an independent reliability ground truth.

core and emo: two specified aggregations. The equal-weight composite and preference-emphasized composite are

$$\mathrm{core}_a = \frac{\mathrm{EF}_a + \mathrm{GD}_a + \mathrm{GR}_a}{3}, \qquad \mathrm{emo}_a = \lambda_{\mathrm{EF}}\mathrm{EF}_a + \lambda_{\mathrm{GD}}\mathrm{GD}_a + \lambda_{\mathrm{GR}}\mathrm{GR}_a. \tag{35}$$

Here the nonnegative coefficients sum to one; our implementation uses $(\lambda_{\mathrm{EF}}, \lambda_{\mathrm{GD}}, \lambda_{\mathrm{GR}}) = (0.5, 0.25, 0.25)$. Full EmoReward scores 0.922 on core and 0.884 on emo, leading under both specified weightings. Reflect without memory leads EF. These observations concern the defined composites and comparison set, not arbitrary weightings or universal superiority.

H Extended Evaluation and Ablations

Provider-pool construction. Figure 5 presents the main-text visual comparison; Table 5 quantifies the same four reward configurations used to construct the provider interface. Semantic emotion and motion cues improve some axes while exposing trade-offs on text alignment and preference. The configurations also differ in fusion weights and curricula; their comparison is not a single-factor ablation. In the semantic-cue configuration, the CLIP channel has zero fusion weight and cannot explain a training gain attributed to an active CLIP reward.

Table 5: **Provider-pool configurations.** Higher is better. TA denotes text alignment. The reward configurations use different weights and curricula; missing VQ is not imputed.

Configuration	Overall	VQ	MQ	TA	ImageReward
VideoAlign	-1.614	-0.505	-0.313	-0.797	-0.866
+Semantic emotion	-1.363	—	-0.014	-0.991	-0.922
+Motion naturalness	-1.453	-0.505	0.075	-1.024	-1.293
Balanced tool pool	-1.178	-0.297	0.074	-0.954	-1.113

Ranking agreement. For each four-candidate group, let $\mathcal{C} = \{(i,j) : i < j,\ y_i \neq y_j\}$. We use

$$\mathrm{Agr}(\mathcal{G}) = \frac{1}{|\mathcal{C}|} \sum_{(i,j)\in\mathcal{C}} \left[\mathbb{I}\{(R_i - R_j)(y_i - y_j) > 0\} + \tfrac{1}{2}\mathbb{I}\{R_i = R_j\}\right]. \tag{36}$$

Groups without comparable pairs are excluded. Agreement is averaged over groups; reward ties receive half credit independently of sample ordering. The analysis contains 160 baseline groups for DanceGRPO and 168 groups for each other optimizer–reward configuration, totaling 1,504 groups and 6,016 videos. These training candidates are separate from the 30-prompt fixed-checkpoint evaluation sets.

Memory sensitivity. Across 27 temperature, neighborhood-size, and memory-window settings, the memory rule exceeds fixed weighting at the final window in 23 settings. None has a positive difference when averaged over all windows. These settings reuse the same traces and are not independent training seeds. The results motivate sensitivity analysis rather than a claim of monotonic improvement with accumulating memory.

I RewardMemory: A Worked Rollout Example

Figure 7 connects recorded provider scores, a previously exported calibration table, and a reproducible allocation calculation. The clips contain 81 frames at 24 fps and have 624×352 resolution; we display frames 12, 40, and 72. The four videos are candidates for one prompt from a reference rollout group. Their scores and outcomes come directly from the accompanying reward records; the images are decoded full frames. The macro provider has positive within-group outcome correlation, while three channels return their neutral constant. These four examples illustrate local behavior, not the population estimate used to construct RewardMemory.

Table 6: **RewardMemory calibration.** Provider reliability and unnormalized multipliers, estimated from 3,232 reflective reward traces ($\tau = 0.5$, $c = 1$). Macro includes VideoAlign information, so its association is partly shared with the outcome.

Provider	Macro	Motion	CLIP	R1-Omni	Micro	Body	Temporal
ρ_k	0.429	0.332	0.073	0.047	0.000	0.000	0.000
m_k	1.833	1.510	0.899	0.854	0.778	0.778	0.778

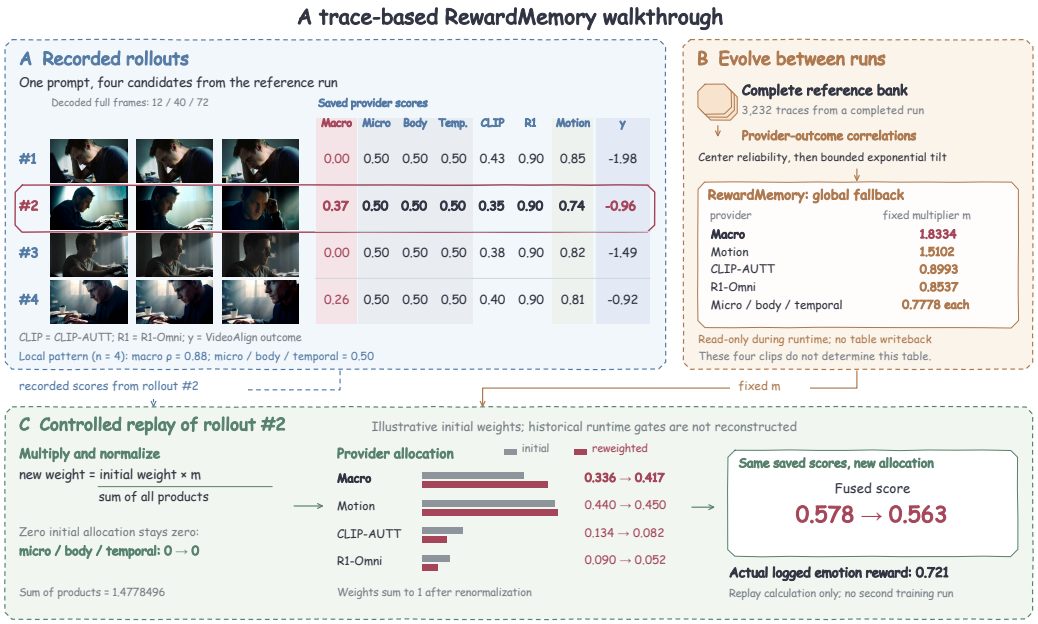


Figure 7: **From completed-run evidence to calibrated fusion.** Recorded observations (A) supply provider scores and outcomes; RewardMemory (B) supplies fixed multipliers estimated from the complete reference bank, rather than the four displayed clips. A controlled replay (C) applies these multipliers to illustrative initial weights for rollout #2. The resulting $0.578 \rightarrow 0.563$ calculation demonstrates reallocation with fixed scores; it is distinct from the logged reward 0.721 and does not represent a second training run.

Separating observations from the replay. The exported table's global entry differs from its context-specific entries. We use the global entry here to isolate the reweighting operation, without claiming that it was the entry selected for these historical videos. The initial active allocation is $(0.336, 0.440, 0.134, 0.090)$ over macro, motion, CLIP-AUTT, and R1-Omni, with zero mass on the other three providers. These are illustrative starting weights, not reconstructed runtime gates. Multiplying by the exported global multipliers and renormalizing produces approximately $(0.417, 0.450, 0.082, 0.052)$. On rollout #2's recorded scores, the weighted mean changes from 0.578 to 0.563; its actual logged emotion reward is 0.721. The distinction matters because the historical reward used its own training phase and configuration. An increase in a provider's allocation need not increase the scalar fused score, and neither change alone establishes improved generated quality.

Prompt. *The camera slowly pushes forward, a cold side light slanting through the window frame, carving a sharp beam across the dimly lit study and illuminating the man seated at his desk. He leans slightly forward, hands folded over the keyboard, fingers tapping the keys with slow, precise rhythm. Shot from a low angle behind and to the side, the frame captures his furrowed brow and narrowed eyes, his gaze fixed like a needle on the flickering stream of code on the screen. His lips are pressed into a thin line, the jawline taut with concentration, nostrils fluttering faintly with each breath, occasionally biting his lower lip as if swallowing an unspoken question. The light casts a soft shadow across his forehead, making the sharp glint in his eyes even more intense. In the background, scattered papers and empty coffee cups rest silently on the bookshelf, while the rhythmic*

tapping of the keyboard echoes in the stillness, like the pulse of a heartbeat. He suddenly pauses, raises his right hand to his temple, his eyes briefly drifting, then swiftly refocuses on the screen—a quiet resolve, a steadfast defiance against time.

J GENERALIZATION TO IMAGE GENERATION

We evaluate Stable Diffusion 3.5 and FLUX with three optimizers and three reward configurations on 256 DrawBench prompts per checkpoint: 18 settings and 84 checkpoints. Internal Emo measures the emotion-stack reward; HPSv3 is a general preference objective also used to train the baseline. Neither independently measures perceived emotion.

SD3.5 baseline results average three checkpoints at 420–500 updates; calibrated variants average five at 2,780–3,000. FLUX baseline checkpoints span 2,760–3,000, and calibrated checkpoints 2,780–3,000. Standard errors describe checkpoint variation, not independent training seeds. Per-prompt arrays support the reported means; paired significance requires complete sample and generation-seed mappings and is not claimed.

Observed transfer pattern. With SD3.5–DanceGRPO, HPSv3 rises from 9.552 (baseline) to 10.096 (Reflect) and 10.302 (EmoReward). SD3.5–FlowGRPO declines on both HPSv3 and internal Emo; FLUX's HPSv3-optimized baseline leads on HPSv3 across optimizers. Thus transfer is objective-dependent rather than uniformly beneficial.

J.1 QUALITATIVE COMPARISON PROTOCOL

Image plates use four curated prompts per displayed setting; video plates cover three optimizers. Figure 5 separately compares provider configurations. These examples illustrate behavior, not a randomly estimated win rate.

Images retain all original pixels. SD3.5 baseline outputs use 500 updates and calibrated outputs 3,000. The first two FLUX baseline cases use the only saved examples at 500 updates; all other FLUX rows use 3,000. Unequal budgets prevent attributing qualitative differences solely to calibration. Image seeds are unavailable, so only prompts are matched. Video rows share prompts and recorded sampling seeds. Clips are 864×480 at 24 fps; full frames at 0.5, 2.5, and 4.5 seconds are decoded without interpolation or retouching. The source bundle includes synchronized playback and editable figures.

The ordinal ratings in Appendix B complement these proxy-based and qualitative comparisons; raw participant/prompt records are needed to quantify uncertainty. The depicted scenes are generated content.

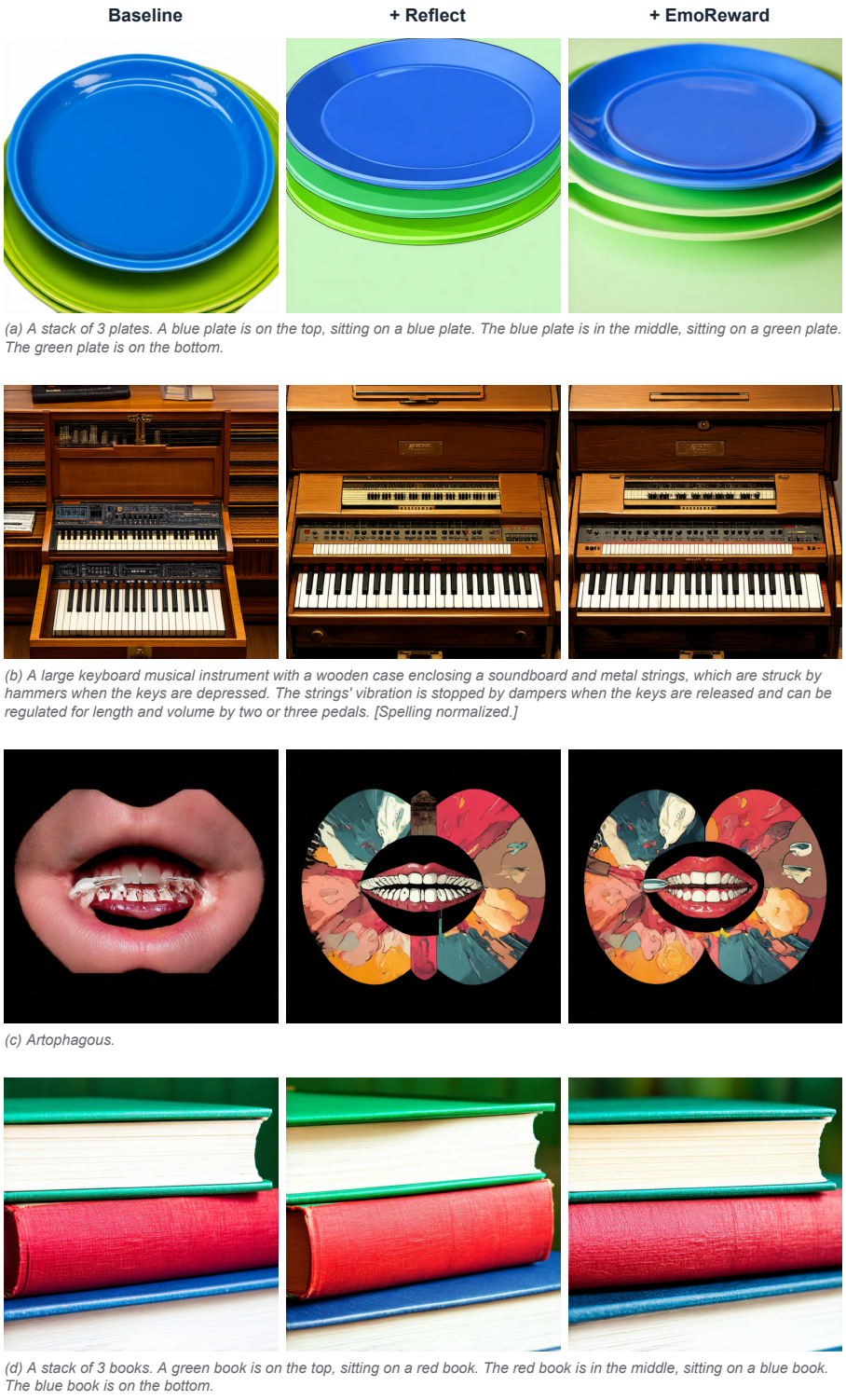
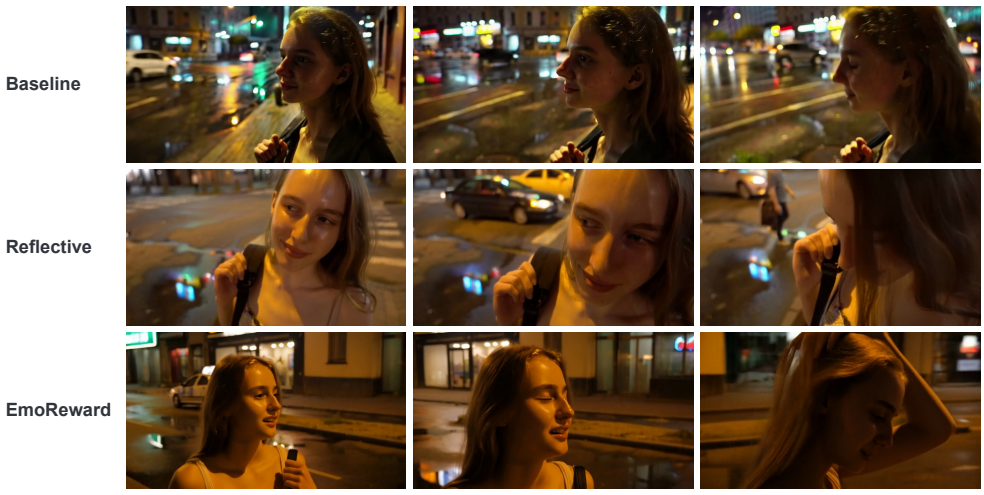


Figure 8: **Qualitative comparison of reward calibration.** Columns show the baseline, the addition of Reflect, and the complete EmoReward configuration for SD3.5 with DanceGRPO. Prompt (b) is spelling-normalized for readability; the images retain the original perturbed input. Budgets and seed limitations appear in Appendix J.1.



Figure 9: Image transfer across SD3.5 and FLUX. Five backbone–optimizer settings share four prompts; each panel compares HPSv3 RL, Reflective gating, and EmoReward. Rows show the same curated prompts as the preceding figure. The first two FLUX baseline rows use earlier saved outputs; see Appendix J.1 for budgets and seed limitations.

K VIDEO COMPARISONS



Prompt: The camera slowly pushes in on a young woman standing at a street corner under the dim yellow glow of a streetlight, captured in a medium close-up. Side backlighting outlines her slender silhouette. Her hand unconsciously tightens around the backpack strap, knuckles turning slightly white, while a faint, forced smile tugs at her lips—never reaching her eyes. Her gaze drifts away from the front, settling on a puddle on the ground, where neon signs flicker and reflect like the instability of her emotions. She slightly turns her head, her lips twitching as if about to speak, but only a soft cough escapes. She lowers her head, her long hair falling to veil half her face. A breeze lifts her strands, and she instinctively reaches up to brush them aside, her fingers pausing in midair before dropping weakly. In the background, blurred traffic and pedestrians move past, their noise isolated beyond the frame, leaving her alone in the shifting light, trembling slightly, as if the entire world has gone silent except for the clear, pounding rhythm of her heartbeat in her ears.

Figure 10: **DanceGRPO.** Street-corner hesitation: a strained smile and averted gaze.



Prompt: The slanting sunlight of dusk filters through the old window lattice, bathing the room in a warm, amber glow. An middle-aged couple sits on worn wicker chairs; she is bent over, mending an old sweater, her fingers deftly guiding the needle and thread. He watches her quietly, his fingers unconsciously tracing the rim of his cup, his gaze reflecting the swaying tree shadows outside. Suddenly, she looks up at him, a faint smile lifting the corners of her lips, fine lines crinkling at the edges of her eyes—gentle ripples of time's quiet grace. His throat moves slightly, his lips parting as if to speak, but no sound comes. Instead, he reaches out, gently brushing a strand of her hair, his touch slow and reverent. A hint of shyness flickers in her eyes; she lowers her head with a soft laugh, her fingers trembling slightly as a thread slips free. He catches it with a quick, tender motion. They exchange a glance, and a silent understanding flows between them. The camera slowly pushes in, focusing on the delicate smile lines at the corner of her eye and the quiet glimmer deep in his gaze.

Figure 11: **DanceGRPO.** A quiet exchange at dusk: the prompt couples a smile with a tentative hand gesture.



Prompt: The camera slowly pushes in on a young woman kneeling on ancient bluish stone steps, her back turned to the deepening twilight sky, hands folded gently on her knees, fingertips trembling faintly. Her face is bathed in warm gold by the slanting rays of the setting sun, her eyelashes lowered, her pupils reflecting the flickering oil lamp at the distant temple eaves—a silent mirror to some unspoken obsession. Her lips are slightly pursed, the line of her mouth taut, as if silently praying or suppressing a sigh that threatens to break free. She tilts her head upward, gazing into the night sky, a single tear slipping down her cheek, pausing at her chin, glimmering in the dim light. A breeze stirs her hair, her sleeve lifts slightly, her arm quivers faintly, as though awaiting a response that will never come. Behind her, a quiet courtyard lies still, wild grasses growing beside the stone steps, while in the distance, a bell tolls gently, echoing through the empty dusk. The entire scene is enveloped in a near-sacred sense of solitude.

Figure 12: **FlowGRPO.** Unease on the stone steps at twilight.



Prompt: The camera slowly pushes in on a young woman sitting alone in the corner of an old-fashioned café, rain streaking diagonally across the window, blurring the street scene through water-traced glass. She holds a cup of coffee now cold, her fingertips trembling slightly, her gaze distant as she stares at the table, lips pressed into a thin line, as if chewing on some unspeakable contradiction. Light slants through the high window, bathing one side of her face in warm yellow while the other remains in shadow, creating a stark visual divide. Suddenly, she lifts her eyes, scanning the empty seats before returning her gaze to her phone screen, her brow furrowing slightly, her mouth twitching as if she's heard a whisper she shouldn't have. She wipes her eye with the back of her hand—restrained yet unable to hide her weariness—her fingers unconsciously picking at the rim of the cup, as if searching for some confirmation that has long since vanished. The air is thick with the scent of coffee and damp wood; time seems to have stopped, except for the flicker of doubt in her brow, quietly growing in the silence.

Figure 13: **FlowGRPO.** A pause in a rain-streaked café.



Prompt: The camera slowly pushes in, revealing a middle-aged woman standing at the center of the stage, bathed in a spotlight that cascades from the upper side, softly outlining her silhouette with quiet strength. She tilts her head slightly upward, a subtle smile forming on her lips—a smile long held within, unassuming yet powerful enough to pierce the silence. Her gaze sweeps across the audience, her eyes glowing with a gentle warmth, like melting snow in winter, reflecting countless looks of reverence. Her fingers gently trace the emblem on her chest, her fingertips trembling slightly, as if touching a memory sealed by time. Slowly, she raises her hand and grasps the microphone, her lips parting slightly as her voice emerges—low, clear, each word like amber seeping from the depths of her soul. Her eyes never leave the audience, a look of calm acceptance and quiet response, as if saying: every step I've taken is worth being seen. Behind her, the applause blurs into soft waves, the melody gentle and hushed, while her face, stretched by light and shadow, resembles a still oil painting, frozen in her moment of radiant glory.

Figure 14: **SAGE-GRPO.** Composure under the spotlight in a restrained, emotionally charged stage scene.



Prompt: The camera slowly pushes in, settling on a close-up of a middle-aged woman seated in an old wicker chair by the window, sunlight slanting across her slightly weary face, fine lines around her eyes trembling faintly in the light. She gazes down at an old letter in her hands, lips slightly parted, breath catching, as if reading a memory sealed by time. Suddenly, a breeze stirs, gently turning the page. She lifts her eyes, her gaze falling on a cherry blossom tree blooming outside, a flicker of recognition passing through her, a soft smile gradually forming, her brow relaxing into a long-forgotten tenderness. She raises a hand, fingertips brushing lightly over the letter, as if touching a distant moment; tears well in her eyes but do not fall, only a faint glimmer shining deep within her pupils. The scene slowly pulls back, revealing her sitting in a room bathed in shifting light and shadow, the flower shadows outside swaying gently, the entire space filled with a quiet, warm aura of bygone days.

Figure 15: **SAGE-GRPO.** Recognition in an old letter in a close-up with gradually softening expression.