STAR: SELECTIVE TIMESTEP ANCHOR REWARD FOR IDENTITY-PRESERVING VIDEO GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: STAR clearly improves identity preservation. Each row shows the first frame, then Base and STAR at two matched times, for single-person (rows 1–2) and multi-person (rows 3–4) scenes. Under large pose and expression changes, Base faces drift from the first frame, while STAR keeps every subject recognizable.

ABSTRACT

Identity-preserving image-to-video post-training must strengthen facial identity without unnecessarily suppressing motion. We introduce STAR, which addresses this trade-off by selectively placing reward gradients along the denoising trajectory. It calibrates an anchor window from the joint response of identity similarity and facial dynamics to timestep-wise perturbations. Within a single-leap backpropagation framework, it samples anchors from this window while an identity-forward connector evaluates the reward on the fully generated video. The resulting update separates timestep coverage from backpropagation depth. Our analysis characterizes terminal scoring, endpoint attenuation of raw reward gradients, and the local identity–motion trade-off of anchor selection. We build HQ4000, a 4,000-clip dataset with per-subject pose and expression annotations, and adapt an identity verifier to support training. We evaluate STAR on four I2V settings: HunyuanVideo-1.5 with 50 and 12 sampling steps, Wan2.2 with 50 steps, and MiniMax-H3. STAR improves test500 identity similarity by 9.8%–35.6% over the base models, with facial dynamics well preserved and no visible reward hacking. Our project page is available at `https://anonymous.4open.science/w/STAR-2027/`.

1 INTRODUCTION

Diffusion Transformers (Peebles & Xie, 2023) have become a foundation for video generation, including HunyuanVideo (Kong et al., 2024; Tencent Hunyuan Foundation Model Team, 2025),

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.

Wan (Team Wan et al., 2025; Team Wan, 2025), and MiniMax-H3 (MiniMax, 2026). In image-to-video (I2V) generation, visual quality alone is insufficient: the generated person should remain recognizable as the reference subject throughout changes in pose and expression (Figure 1). Reward post-training offers a direct way to optimize this requirement, but can also suppress facial movement or favor easier, reference-aligned appearances.

Efficient differentiable post-training uses truncated or leap-based gradients instead of retaining the full denoising graph. This leaves an important design choice: which sampling timesteps should carry the identity update? A short continuous suffix controls backpropagation depth but also restricts its temporal coverage. A randomly selected leap can reach more of the trajectory, yet the selected timesteps need not offer the same identity benefit or motion cost. The central issue is therefore not only how far to backpropagate, but where to place the available gradient computation.

Two limitations motivate this design: a memory-constrained gradient horizon and a reward interface requiring task-specific adaptation. IPRO-style continuous backpropagation becomes more expensive as its horizon grows. Identity-GRPO uses a preference-trained VLM, whereas DreamVideo-Omni scores features before VAE decoding. STAR combines a common RGB scoring interface with a calibrated identity–dynamics anchor window; gradients still pass through each generator's own decoder.

We introduce STAR, *Selective Timestep Anchor Reward*. A timestep probe measures how a one-step clean prediction responds to an identity-directed perturbation, jointly tracking identity gain and facial-dynamics change. These responses identify a backbone-specific region with sustained identity sensitivity and limited dynamic disturbance. STAR samples one anchor from this region, recomputes its prediction with gradients, and scores the true terminal video through an identity-forward connector. The probe selects the update location; terminal scoring keeps the reward attached to the actual generated output.

Our contributions are threefold:

1. **Probe-guided anchor selection.** A timestep probe measures how an identity-directed perturbation at each sampling step changes identity similarity and facial dynamics. After a scheduler pre-screen on $q(k)$ removes highly compressed leaps, we take the interval with sustained identity response and small dynamic disturbance as the STAR window for anchor sampling.
2. **Identity-consistency training.** STAR separates timestep coverage from backpropagation depth: each update backpropagates through a single anchor from this window, and the retained graph stays one DiT step wherever the anchor lies, while an identity-forward connector scores the fully generated video. A teacher-velocity penalty at the same anchor regularizes the update. On four I2V settings, STAR improves test500 identity similarity by 9.8%–35.6% over the base models.
3. **Supporting dataset and verifier.** HQ4000 provides 4,000 single- and multi-person clips with per-subject, pose/expression-aware reference pools. The identity reward is the minimum over three ArcFace-R100 verifiers fine-tuned with generated identity-drift frames as hard negatives (min3). Construction and training details are given in Appendices C and D.

2 Related Work

Identity-preserving generation. Phantom (Liu et al., 2025b), MAGREF (Deng et al., 2025), HunyuanCustom (Hu et al., 2025), and Kaleido (Zhang et al., 2025) preserve subject identity through reference-conditioned generation. For face-centric identity, ConsisID (Yuan et al., 2025b) decomposes the reference face into low- and high-frequency signals and injects them at different depths of a DiT, while Concat-ID (Zhong et al., 2025) concatenates VAE-encoded reference latents along the sequence dimension and lets the existing 3D self-attention perform the fusion without extra modules. These methods modify the conditioning pathway at inference time and are complementary to post-training an existing generator with an identity reward, which is our setting.

Differentiable reward post-training. DRaFT-K (Clark et al., 2024) backpropagates through a continuous suffix of K sampling steps; ReFL (Xu et al., 2023) uses a late-step prediction. DRTune (Wu et al., 2024) studies deep reward supervision using stopped denoiser-input gradients and sampled steps. LeapAlign (Liang et al., 2026) uses shortened leap trajectories and latent connectors to propagate terminal rewards. STAR builds on this efficient-gradient family and focuses on calibrating an identity–dynamics anchor window, keeping the backpropagation depth at a single step.

Identity reward post-training. IPRO (Shen et al., 2026) applies a differentiable ArcFace reward through a continuous denoising suffix. Its backpropagation length K is constrained by the memory/computation budget: reaching earlier timesteps requires a longer retained or recomputed chain. The constraint applies to the gradient horizon, while the number of forward sampling steps stays unaffected. Identity-GRPO (Meng et al., 2025) learns a VLM identity reward from preferences and uses group-relative optimization. Its reward must remain reliable on the target generator's output distribution, and each update compares multiple sampled candidates. DreamVideo-Omni (Wei et al., 2026) learns an identity reward on intermediate video latents, bypassing VAE decoding and sampling feedback across diffusion timesteps. This saves a decoder pass in the reward path, but ties the input interface to the VAE latent representation. Transferring a reward trained in one VAE's coordinates to an incompatible VAE's latents requires alignment, re-encoding, or retraining. Table 7 provides the method summary.

3 METHOD

3.1 PROBLEM SETUP AND IDENTITY OBJECTIVE

Let v_θ be an I2V generator whose N-step sampler produces a terminal latent x_N. For a subject's valid generated frames $\mathcal{F}$, let $\hat{g}_{fm}$ and $\hat{r}_{jm}$ be unit face embeddings from verifier member $m \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. A pose-aware reference pool defines

$$R_{\mathrm{id}} = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \min_m \cos(\hat{g}_{fm}, \tilde{r}_{fm}), \qquad \tilde{r}_{fm} = \sum_j \alpha_{fj} \hat{r}_{jm}. \tag{1}$$

The cosine normalizes the weighted prototype; min3 is taken *before* averaging frames. For multiple subjects we average their rewards. Reference weights are

$$\alpha_{fj} = \frac{\exp(-\mathrm{dist}_{fj})}{\sum_{j'} \exp(-\mathrm{dist}_{fj'})}, \quad \mathrm{dist}_{fj} = \frac{|\psi_f - \psi_j|}{\tau_\psi} + \frac{|\rho_f - \rho_j|}{\tau_\rho} + \frac{\|e_f - e_j\|_2}{\tau_e}, \tag{2}$$

with yaw ψ, pitch ρ, and a 212-dimensional expression descriptor e; $(\tau_\psi, \tau_\rho, \tau_e) = (20^\circ, 15^\circ, 0.15)$. Detection, subject assignment, and reference weights are stopped on the backward branch; gradients pass through face crops and verifier embeddings (Appendix A).

We optimize identity reward with a teacher-velocity penalty,

$$\mathcal{J}(\theta) = \lambda \, \mathbb{E}_\xi R_{\mathrm{id}}(x_N(\theta; \xi)) - \beta \, \mathbb{E}_{\xi, t^\star} \mathcal{L}_{\mathrm{vel}}(\theta; \bar{x}_{t^\star}), \tag{3}$$

where ξ contains conditioning and sampling randomness, $t^\star \sim \mathcal{U}(\mathcal{W})$, and $\bar{x}_{t^\star}$ is the stopped anchor input. The penalty is the mean squared difference between policy and frozen-teacher velocities at that input. It is a velocity regularizer; the conditions under which it reads as a KL under Gaussian transitions are given in Appendix B. Eq. (3) specifies the scored objective; its single-anchor surrogate gradient is given in §3.3.

3.2 TIMESTEP PROBE AND STAR-WINDOW CALIBRATION

Joint-response probe. At sampling index k (zero-based), the stopped latent and velocity define a one-step clean prediction. We perturb the velocity by $\Delta v = \varepsilon \|v_k\|_2 \hat{\gamma}_{\mathrm{id}}$, where $\hat{\gamma}_{\mathrm{id}}$ is the unit gradient of its identity reward with respect to velocity, then decode the perturbed prediction. We measure identity gain and the absolute change in temporal standard deviations:

$$\Delta\mathrm{Dyn} = |\Delta\mathrm{std}_{\mathrm{yaw}}| + w_p |\Delta\mathrm{std}_{\mathrm{pitch}}| + w_e |\Delta\mathrm{std}_{\mathrm{expr}}|. \tag{4}$$

The weights w_p, w_e match RMS scales on a fixed calibration set (Appendix B). $\Delta\mathrm{Dyn}$ measures a local *probe response* to a single perturbation; the motion retained after full post-training is reported as Dyn$\times$ in §5.6.

Scheduler pre-screen. Define $q(k) = \sigma_k / (\sigma_k - \sigma_{k+1})$ from consecutive sampler noise levels; it counts how many local steps the one-step leap from index k compresses into one extrapolation. The probe scores this one-step prediction, a faithful surrogate of the final video only for short leaps: for large $q(k)$ it is blurry, face scoring on it is unreliable, and its gap to the sampled terminal, which

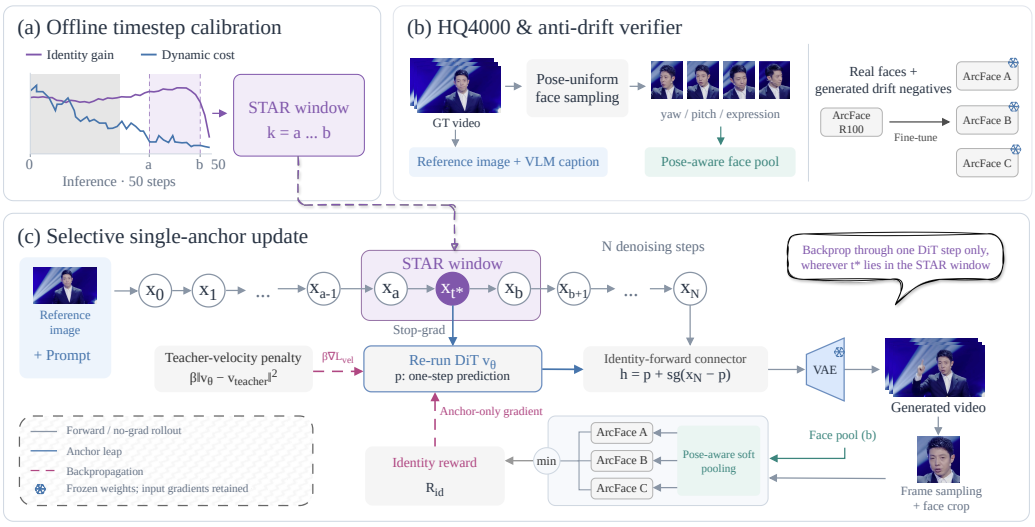


Figure 2: Overview of STAR. (a) Joint identity–dynamics probing calibrates the anchor window. (b) HQ4000 and the adapted verifier support pose-aware scoring. (c) A no-gradient rollout supplies the terminal; one anchor carries gradients through the connector, with teacher-velocity regularization at the same anchor.

bounds the proxy error in Eq. (6), grows. We therefore retain $q(k) \leq 45$, a threshold set by empirical admissibility on the calibration set.

Window calibration. Within the admissible region, the probe guides an interval with sustained identity response and reduced dynamic disturbance, excluding the terminal response collapse; the resulting windows are listed in Appendix B, and a reference recalibration interface selects a contiguous feasible interval under explicit identity and disturbance thresholds chosen on calibration data. Training samples anchors uniformly from the fixed window.

3.3 Selective Single-Anchor Update

Figure 2 summarizes the procedure. A full no-gradient rollout first obtains the true terminal x_N and the stopped anchor input. We then recompute only the selected anchor with gradients and form its one-step terminal prediction $p_{t^\star}$. The identity-forward connector is

$$h_{t^\star} = p_{t^\star} + \mathrm{sg}(x_N - p_{t^\star}), \tag{5}$$

where sg denotes stop-gradient. Its forward value is x_N, while its backward path passes only through $p_{t^\star}$. Thus the reward is evaluated on the generated video with a single retained DiT step, and the policy is updated along the single-anchor surrogate gradient $\lambda(\partial_\theta p_{t^\star})^\top \nabla R_{\mathrm{id}}(x_N) - \beta \nabla_\theta \mathcal{L}_{\mathrm{vel}}(\theta; \bar{x}_{t^\star})$ of Eq. (3). The retained backpropagation depth does not grow with the anchor window or sampling-step count; the full forward rollout is still required. This construction uses the latent-connector principle of LeapAlign (Liang et al., 2026) with a selectively sampled single anchor.

The window and teacher penalty jointly set the observed operating point, whose motion preservation is measured in §5.6.

3.4 Reference Data and Adapted Identity Scoring

HQ4000. We build 4,000 training clips: 2,893 single-person and 1,107 multi-person scenes. The dataset covers 5,115 supervised subjects and over 125k face crops with yaw/pitch, 212-dimensional expression features, and identity embeddings. Per-subject reference pools support pose-aware scoring and multi-person reward matching. Captions are generated by a locally deployed Qwen3.5-397B (Qwen Team, 2026). The full construction pipeline is given in Appendix D.

Algorithm 1 STAR

Input: policy v_θ, frozen teacher, identity verifier, λ, β.
Calibration: probe each timestep; measure identity and dynamic responses of the clean prediction; select $\mathcal{W}$.
Each update:

1. Sample $t^\star \sim \mathcal{U}(\mathcal{W})$; run the full sampler without gradients, caching $\bar{x}_{t^\star}$ and x_N.
2. Recompute the anchor with gradients and form $p_{t^\star}$.
3. Set $h_{t^\star} = p_{t^\star} + \mathrm{sg}(x_N - p_{t^\star})$; decode and compute the min3 identity reward.
4. Compute the teacher-velocity penalty at the same stopped anchor; update the policy along the surrogate gradient.

Adapted verifier. We fine-tune public ArcFace-R100 (Deng et al., 2019; InsightFace Contributors, 2026) using real faces and generated hard negatives. The `drift_neg` term includes generated drift frames as per-identity negatives. Three complementary configurations yield member scores $s_{\mathrm{A}}, s_{\mathrm{B}}, s_{\mathrm{C}}$; their minimum, termed *min3*, supplies the frame reward. This conservative aggregation follows reward-ensemble practice (Coste et al., 2024). Gradients flow through the minimizing member on a locally fixed active branch. Reference faces are softly weighted by pose/expression proximity as in Eq. (2). Offline evaluation appears in §5.5; verifier losses are in Appendix C.

4 ANALYSIS OF THE STAR DESIGN

Let $F(\theta) = \mathbb{E}_\xi R_\xi(X_\theta)$ be full-resampling identity reward, with decoding included in R_ξ, and write $A_t = \partial_\theta p_t$, $J = D_\theta X_\theta$, $\omega = \nabla R_\xi(X_\theta)$. Appendix A gives the derivations and assumptions.

Why score the sampled terminal? With faithful differentiation on the reward branch, the connector gives $\gamma_t = A_t^\top \omega$, whereas direct leap scoring gives $\gamma_t^{\mathrm{proxy}} = A_t^\top \nabla R_\xi(p_t)$. For an L_R-Lipschitz reward gradient between p_t and X_θ,

$$\begin{aligned} \gamma_t - J^\top \omega &= (A_t - J)^\top \omega, \\ \|\gamma_t^{\mathrm{proxy}} - \gamma_t\| &\le \|A_t\| L_R \|p_t - X_\theta\|. \end{aligned} \tag{6}$$

Terminal scoring removes the scoring-location discrepancy, not the single-anchor Jacobian bias.

Why stop before the sampling endpoint? For a fixed cached input, write the one-step extrapolation as $p_t = \bar{x}_t + c_t v_\theta(\bar{x}_t, t)$, where c_t is the remaining signed integration coefficient, independent of θ. If $\|\partial_\theta v_\theta\| \le K_v$ and $\|\omega\| \le K_R$ uniformly over the anchors considered, then

$$A_t = c_t \partial_\theta v_\theta, \qquad \|\gamma_t\| \le |c_t| K_v K_R \longrightarrow 0 \quad \text{as } c_t \to 0. \tag{7}$$

Raw reward gradients attenuate as the extrapolation length vanishes, so the window must also satisfy a lower bound on identity response. This motivates an upper boundary; the finite cutoff is set by the probe curves. The attenuation applies to the raw reward contribution, while the AdamW update magnitude is governed separately by preconditioning and momentum.

Why calibrate identity and dynamics jointly? For actual optimizer directions $\widehat{d}_t$ including regularization, let $a_t = \langle \nabla F, \mathbb{E}\widehat{d}_t \rangle$, $M_t = \mathbb{E}\|\widehat{d}_t\|^2$, and b_t be the expected first-order motion-descriptor disturbance. For $\eta > 0$, uniform anchors, window averages denoted by overbars, local smoothness, and a uniformly bounded descriptor Taylor remainder,

$$\begin{aligned} \mathbb{E}[F(\theta + \eta\widehat{d}_t) - F(\theta)] &\ge \eta \bar{a}_{\mathcal{W}} - \tfrac{1}{2} L \eta^2 \bar{M}_{\mathcal{W}}, \\ C_{\mathcal{W}} &\le \eta \bar{b}_{\mathcal{W}} + \tfrac{1}{2} H \eta^2 \bar{M}_{\mathcal{W}}. \end{aligned} \tag{8}$$

Here $C_{\mathcal{W}}$ measures expected full-resampling motion-descriptor disturbance; Appendix A defines its norm and L, H. The bounds favor high identity response, low disturbance, and controlled update magnitude. The clean-prediction probe guides this tradeoff empirically; direct measurement of a_t, b_t and a proof of window optimality are left to future work.

Table 1: Main results (base vs. STAR): test500 min3 FaceSim; VBench-I2V official scores in %; OpenS2V Aes./Gme/Mot_{fb}/Smo.; VBench 2.0 Clothes.

		test500↑	VBench-I2V ↑									OpenS2V ↑				VB2↑
Model		FS	Sub.	Bg.	SC	BC	Mot.	Dyn.	Flk.	Aes.	Img.	Aes.	Gme	Mot	Smo.	Clo.
HY1.5-50	base	0.5793	98.60	99.37	95.11	97.02	99.19	52.44	98.05	64.62	70.17	5.064	0.5234	0.3852	0.9238	0.8169
	STAR	**0.7249**	98.60	99.39	95.46	97.17	99.20	49.59	98.13	64.77	70.30	5.125	0.5847	0.3932	0.9405	0.8750
MiniMax-H3	base	0.6545	98.22	99.08	95.16	97.10	99.08	50.00	97.86	61.60	70.04	5.468	0.5645	0.3150	0.9533	0.9041
	STAR	**0.7183**	98.20	99.05	95.13	96.93	99.08	52.03	97.79	61.52	70.07	5.467	0.5591	0.2976	0.9592	0.9437
Wan2.2-50	base	0.5379	98.50	99.43	95.28	97.18	98.62	52.03	97.38	64.19	70.97	5.272	0.5143	0.3970	0.9035	0.7867
	STAR	**0.7217**	98.41	99.34	95.17	96.49	98.64	51.63	97.45	63.35	70.95	5.264	0.5087	0.3976	0.9052	0.8400
HY1.5-12	base	0.5670	98.78	99.49	95.81	97.78	99.32	43.09	98.51	65.35	70.20	5.114	0.5845	0.3309	0.9123	0.8108
	STAR	**0.7691**	98.45	99.39	95.08	97.45	99.15	46.75	98.37	64.85	69.67	5.091	0.5569	0.3168	0.9223	0.8378

5 EXPERIMENTS

5.1 SETUP

Backbones. Our main backbone is HunyuanVideo-1.5 I2V 8B (Tencent Hunyuan Foundation Model Team, 2025) (50-step, CFG-free). To test cross-backbone generality we also experiment on Wan2.2 5B (Team Wan, 2025) (50-step), MiniMax-H3 (MiniMax, 2026) (50-step, CFG-distilled), and the distilled HunyuanVideo-1.5 (12-step); all models are fully fine-tuned. We additionally apply STAR to VACE (Jiang et al., 2025) 1.3B (R2V) as a beyond-I2V check (§5.2).

Data and evaluation. Training uses HQ4000 (§3.4). We additionally build a test set of 500 clips (*test500*: 357 single-person and 143 multi-person) covering large-pose/large-expression scenes.

Optimization. We use the selective-timestep backpropagation of §3.3. Main experiments run on 128 H20 GPUs with a global batch of 64; AdamW (Loshchilov & Hutter, 2019), learning rate 2×10^{-5}. The reward weight is λ=0.1; the velocity-penalty weight β is 10 on HunyuanVideo-1.5 and MiniMax-H3, 5 on Wan2.2 and VACE, and 3 on the 12-step model; the STAR window is calibrated per backbone (Figure 3). The full training configuration is given in Appendix B.

Baselines. We compare against three groups: (i) within the truncated-backprop family—IPRO (continuous K-step backprop, placed on the late steps and inside the calibrated window) and unstructured anchor windows (full-trajectory uniform, first-half, second-half); (ii) a group-sampling RL baseline—a controlled GRPO (Shao et al., 2024; Liu et al., 2025a) reproduction under the same min3 reward; (iii) the learned-RM method Identity-GRPO, testing its cross-backbone transfer (§5.4).

5.2 MAIN RESULTS

We evaluate STAR on the four I2V backbones of §5.1. Beyond test500, three public suites play complementary roles. **VBench-I2V** (Huang et al., 2024) is the native I2V protocol (355 pairs, 480p 121f, 16:9, eight official dimensions plus temporal flickering). **OpenS2V-Nexus** (Yuan et al., 2025a) (Single-Domain, 90 samples, official scripts) reports aesthetics, text–video alignment, motion amplitude, and smoothness under a subject-to-video protocol. **VBench 2.0** (Zheng et al., 2025) Human_Clothes (custom_input) is a clothing-fidelity check—whether identity post-training collapses apparel. Table 1 reports these under third-party extractors.

Two observations. **First**, STAR improves test500 identity versus the reference on every I2V backbone (min3 +0.146 / +0.064 / +0.184 / +0.202 on HY1.5-50, MiniMax-H3, Wan2.2, and HY1.5-12). The gains persist under the raw ArcFace R100 extractor, which is not used as the training reward (+0.139 / +0.075 / +0.199 / +0.168); a single-/multi-person breakdown is in Appendix E. **Second**, VBench-I2V is saturated on subject/background consistency: eight of nine dimensions move by at most 0.9 points under STAR; `dynamic_degree` is the only dimension that moves by more than one point, and not in a consistent direction. OpenS2V aesthetics, Gme, motion and smoothness stay within a few points of the base, and VBench 2.0 clothes does not collapse (+0.058 / +0.040 / +0.053 / +0.027 on HY1.5-50, MiniMax-H3, Wan2.2, and HY1.5-12).

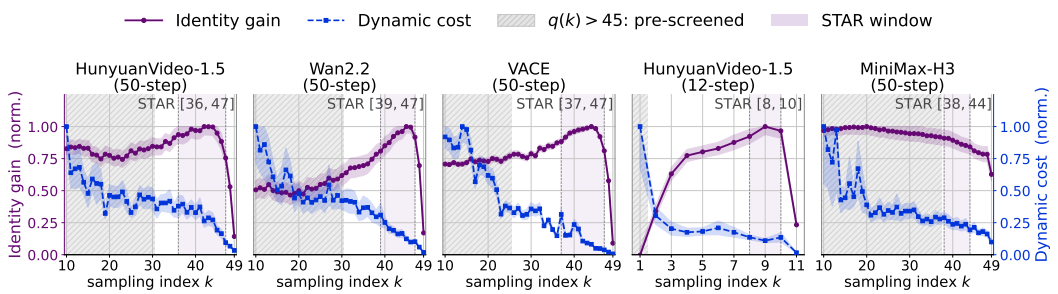


Figure 3: Timestep probe on five backbones; each curve is normalized by its maximum over the plotted steps. Solid: identity gain; dashed: dynamic cost. Gray: $q(k) > 45$; purple: STAR window.

Table 2: Window ablation (HunyuanVideo-1.5, test500).

Window	FaceSim
base	0.5793
$[0, 24]$	0.6780
$[0, 49]$	0.6890
$[25, 49]$	0.7110
STAR $[36, 47]$	**0.7249**

Table 3: Mechanism comparison. Brackets give the backpropagated timestep range.

Method	FaceSim
base	0.5793
IPRO K=4, $[46, 49]$	0.6363
IPRO K=4, $[44, 47]$	0.6594
GRPO	0.5941
Identity-GRPO	0.5467
STAR	**0.7249**

Table 4: Verifier ablation (HunyuanVideo-1.5, test500). Rows: training reward; columns: evaluation metric.

Training reward	raw min3 ↑	raw ArcFace ↑
base	0.5793	0.4569
raw R100	0.6574	0.5325
min3	**0.7249**	**0.5955**

Extra R2V check (VACE). We further apply STAR to VACE 1.3B, an R2V (reference-to-video) backbone: test500 min3 FaceSim rises $0.2044 \rightarrow 0.4651$—the largest identity gain among all backbones—confirming that the selective-anchor recipe transfers beyond I2V.

5.3 Timestep Analysis

The reported probe uses 64 samples per backbone and stride 1, and curves are normalized within panels (Figure 3). The first four settings share one pattern: identity gain peaks late ($k = 42$–45 of 50 steps; $k = 9$ of 12) and collapses in the last one or two steps (to 9%–23% of its peak), while the dynamic cost keeps falling. Their windows cover this late plateau and stop before the collapse, keeping identity gain at 75%–100% of its peak with a mean normalized dynamic cost of 0.13–0.29, versus 0.21–0.39 over the admissible steps before the window.

MiniMax-H3 has a flatter identity curve than the other settings. Its shifted schedule keeps $q(k) > 45$ through $k = 37$, excluding the early plateau under our pre-screen. In its calibrated $[38, 44]$ interval, identity response remains 84–91% of the peak. The window ablation in §5.4 compares the HY1.5 window with coarse alternatives.

5.4 Ablations

We ablate two factors on the HunyuanVideo-1.5 backbone with the test500 protocol—the window position (*which* timesteps to backpropagate through) and the optimization mechanism (*how* to backpropagate)—and additionally ablate the training reward (min3 vs. raw R100).

Window position (Table 2). The calibrated window reaches the highest FaceSim among the tested windows (0.7249 vs. at most 0.7110).

Optimization mechanism (Table 3). With the same identity reward, STAR gains $+0.146$, above the tested IPRO $K = 4$ variants ($+0.057$, $+0.080$) and controlled GRPO ($+0.015$). Identity-GRPO, reproduced with its released reward model, does not improve FaceSim: its reward-model score oscillates upward during training while FaceSim does not follow.

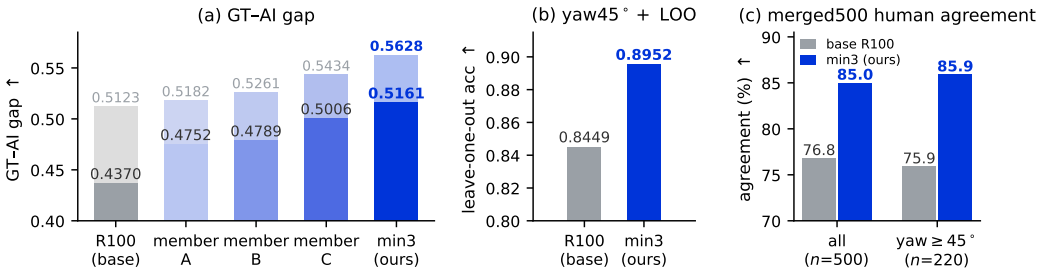


Figure 4: Offline verifier evaluation. (a) GT–AI gap (light bars: soft aggregation) for the original R100, the three fine-tuned members, and min3; (b) yaw$45°+$ leave-one-out same-identity accuracy; (c) verifier–human agreement on merged500.

Table 5: Dynamics and layout (test500). Dyn×, Area×, and Off× are STAR/base (1.00 = unchanged). Dyn× averages yaw/pitch/expression std ratios; Area is largest-face bbox fraction; Off is center offset (lower Off× = more centered).

	Dynamics	Layout	
Model	Dyn×	Area×	Off×
HY1.5	0.89	0.997	0.997
MiniMax-H3	0.88	1.000	0.988
Wan2.2	0.89	0.998	0.991

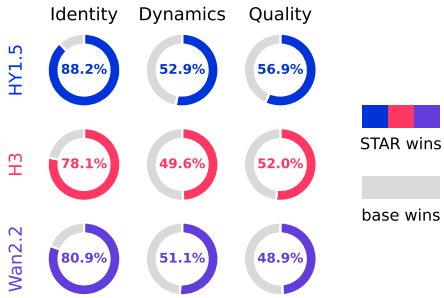


Figure 5: STAR win rates vs. base in the human blind study.

Verifier ablation: min3 vs. raw R100. Holding the STAR window, budget, and data fixed, replacing raw R100 with min3 raises the identity gain from $+0.078$ to $+0.146$ (Table 4)—a nearly twofold improvement due only to the training signal. In this regime the limiter is thus not the selective leap but the reliability of the identity reward on generated frames. Off-the-shelf R100 is poorly calibrated at large pose (HQ4000 clips vary by $>70°$): it assigns spuriously high scores to drifted frames, so the gradient is diverted toward artifacts and the online reward saturates after warmup. min3, a conservative lower bound over three members, suppresses these false positives. The same ranking holds under raw ArcFace, supporting the identity gain under a second evaluation extractor.

5.5 Offline Verifier and Human Evaluation

Verifier evaluation. The verifier uses pose-aware SupCon (Khosla et al., 2020), teacher distillation, and generated drift negatives (Appendix C). On 320 held-out identities / 8.0k crops, Figure 4 reports the real–real minus real–generated similarity gap (GT–AI gap), and large-yaw ($|\text{yaw}| \geq 45°$) leave-one-out matching. Relative to raw R100, min3 raises the gap from 0.4370 to 0.5161 and matching from 0.8449 to 0.8952; soft reference aggregation further increases the gap. The gap reflects both identity-drift sensitivity and real/generated domain discrimination, so we also examine human agreement.

Human evaluation (merged500). We further verify that the verifier aligns with human judgment. Multiple annotators independently judge each item—a real video and two face images—picking "which face looks more like the person in the video," yielding 500 valid items (including a large-pose subset). Taking the human answers as ground truth, we measure the agreement rate between each verifier and the human judgments (Figure 4c).

min3's human agreement exceeds raw R100's by 8.2 points on all items and 10.0 points on the large-yaw subset—the larger the pose, the clearer min3's advantage, consistent with its design goal.

Cost. Each verifier member is fine-tuned for 2,500 steps on one GPU (≈ 0.6 GPU-hours total, $< 0.1\%$ of main training); at inference min3 adds two embedding forwards, below 1% per iteration.

5.6 Dynamics and Reward-Hacking Analysis

Identity-preserving post-training can take shortcuts: suppressing facial motion ("face freezing"), enlarging and centering the face (layout hacking), or overfitting the training verifier. We characterize facial dynamics by the temporal (across-frame) standard deviation of yaw/pitch and expression amplitude, and define **Dyn**$\times$ as the average std ratio vs. the base (1.00 = unchanged). Layout hacking is checked by the STAR/base ratios of largest-face area (**Area**$\times$) and face-center offset (**Off**$\times$; lower = more centered). Table 5 reports both. Training curves are in Appendix F.2.

Identity–motion operating point. Dyn$\times$ is 0.88–0.89 on the three 50-step backbones: measured facial-motion amplitude decreases by 11%–12%. Area$\times$/Off$\times$ remain near one (0.988–1.000): face size and position stay unchanged in aggregate. These two metrics target layout hacking; other shortcuts each require their own test.

On HY1.5, versus scoring the one-step prediction directly, the connector keeps identity similar (0.7249 vs. 0.7271) while raising Dyn$\times$ from 0.85 to 0.89 (Appendix F.3).

Human preferences. In the stratified test500 blind comparison, independent annotators make binary choices on identity, dynamics, and quality (Figure 5). STAR identity win rates are 88.2%, 78.1%, and 80.9% for HY1.5, H3, and Wan2.2. Dynamics rates are 52.9%, 49.6%, and 51.1%; quality rates are 56.9%, 52.0%, and 48.9%. Identity thus improves clearly, while dynamics and quality differ little from the base, with win rates near 50%.

5.7 Qualitative Comparison

Figure 1 pairs Base/STAR frames at matching times. Row 2 uses HY1.5 with a captain on a ship deck; rows 1, 3, and 4 use MiniMax-H3: a person turning beside a white horse, two people scanning a dim hallway, and six people at an awards event. Under these pose and expression changes, the Base faces drift away from the reference, while STAR keeps each subject recognizable with comparable motion, framing, and background. In the multi-person rows, each face is matched to its own subject's reference pool (§3.4), so the update strengthens every identity separately rather than pulling faces toward a single reference. Aggregate statistics are in Tables 1 and 5, and the full videos are on the project page.

6 Discussion and Conclusion

STAR makes the placement of identity gradients along the denoising trajectory a measured choice: a timestep probe calibrates a backbone-specific anchor window, and each update backpropagates through one anchor with a one-step graph while the reward scores the fully generated video. With HQ4000 and the min3 verifier, STAR improves test500 identity by 9.8%–35.6% on four I2V settings and transfers to R2V. On HY1.5, STAR outperforms IPRO and GRPO under the same reward; on the three 50-step backbones, it retains 88%–89% of facial-motion amplitude with face layout unchanged, and human raters clearly prefer its identity while judging dynamics and quality close to the base. Because the retained graph is one DiT step wherever the anchor lies, moving the window earlier costs no extra memory, whereas a continuous suffix such as IPRO's grows with the backpropagation horizon.

The probe is not tied to identity: it needs only a differentiable reward and a descriptor of what the update should preserve. As an extension, with an image-quality, aesthetic, or text–video alignment reward replacing identity, the same procedure could locate the timesteps where that reward responds most and side effects stay small, suggesting a general recipe for placing gradients in reward post-training.

Limitations and future work. The STAR window is shared by all training data of a backbone; adapting its position to each sample is a natural extension. STAR also targets facial identity; consistency of a person's dynamic traits, such as personality and voice, is not addressed.

AI USE STATEMENT

In this work, we used generative AI tools for the following task with required disclosure: data preparation, including generating the HQ4000 training captions with Qwen3.5-397B-A17B (Appendix D). We used generative AI tools to polish the language of the manuscript. AI assistance was not a source of the reported numerical results. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our face data are drawn from an existing internal video library for identity-preserving post-training; identity and pose/expression annotations are produced automatically. Separately, the evaluation includes human preference judgments as described in the experiments. We do not release the underlying face datasets or identity lists. Several qualitative examples in Figure 1 and on the project page start from publicly available images of public figures; they serve only to illustrate identity drift, and every video shown is a synthetic output of the evaluated models. Code and model release must respect the applicable data and model licenses. The use of identifiable faces and human judgments requires attention to consent, privacy, and institutional review requirements; automatic annotation does not remove these responsibilities. Identity-preserving video generation improves the fidelity of synthetic content featuring real people; beyond legitimate scenarios such as film production and personal digital avatars, it may be misused for unauthorized portrait synthesis, impersonation, and the spread of misleading content. We call on downstream use to follow relevant norms on portrait-rights protection and synthetic-content labeling (e.g., watermarks and provenance metadata), and to clearly mark the synthetic nature of generated content when publishing it.

REPRODUCIBILITY STATEMENT

Appendix A gives the derivations and assumptions; Appendix B specifies reported training settings and the revised reference implementation; Appendices C and D describe verifier training and data construction. The accompanying code covers timestep-probe calibration and training for both HunyuanVideo-1.5 and MiniMax-H3. HQ4000 is based on an internal video library and is not released. Matching verifier checkpoints remain a required dependency. Comparable data can be reconstructed with the same pipeline (Appendix D).

REFERENCES

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

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://arxiv.org/abs/1801.07698.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. Magref: Masked guidance for any-reference video generation with subject disentanglement, 2025. URL https://arxiv.org/abs/2505.23742.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation, 2025. URL https://arxiv.org/abs/2505.04512.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models, 2024. URL https://arxiv.org/abs/2411.13503.

InsightFace Contributors. InsightFace: 2D and 3D face analysis project. https://github.com/deepinsight/insightface, 2026. Python library, version 1.0.1.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jiang_VACE_All-in-One_Video_Creation_and_Editing_ICCV_2025_paper.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models, 2024. URL https://arxiv.org/abs/2412.03603.

Zhanhao Liang, Tao Yang, Jie Wu, Chengjian Feng, and Liang Zheng. Leapalign: Post-training flow matching models at any generation step by building two-step trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23238–23248, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liang_LeapAlign_Post-training_Flow_Matching_Models_at_Any_Generation_Step_by_CVPR_2026_paper.html.

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, 2025a. URL https://arxiv.org/abs/2505.05470.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment, 2025b. URL https://arxiv.org/abs/2502.11079.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiangyu Meng, Zixian Zhang, Zhenghao Zhang, Junchao Liao, Long Qin, and Weizhi Wang. Identity-grpo: Optimizing multi-human identity-preserving video generation via reinforcement learning, 2025. URL https://arxiv.org/abs/2510.14256.

MiniMax. MiniMax-H3. Hugging Face model card, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Qwen Team. Qwen3.5-397B-A17B. Hugging Face model card, February 2026. URL https://huggingface.co/Qwen/Qwen3.5-397B-A17B.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Liao Shen, Wentao Jiang, Yiran Zhu, Jiahe Li, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27281–27290, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shen_Identity-Preserving_Image-to-Video_Generation_via_Reward-Guided_Optimization_CVPR_2026_paper.html.

Team Wan. Wan2.2. GitHub repository, July 2025. URL https://github.com/Wan-Video/Wan2.2.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, et al. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Tencent Hunyuan Foundation Model Team. HunyuanVideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Yujie Wei, Xinyu Liu, Shiwei Zhang, Hangjie Yuan, Jinbo Xing, Zhekai Chen, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Ruihang Chu, Yingya Zhang, Yike Guo, Xihui Liu, and Hongming Shan. Dreamvideo-omni: Omni-motion controlled multi-subject video customization with latent identity reinforcement learning, 2026. URL https://arxiv.org/abs/2603.12257.

Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. Deep reward supervisions for tuning text-to-image diffusion models. In *Computer Vision – ECCV 2024*, pp. 108–124. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73010-8_7. URL https://doi.org/10.1007/978-3-031-73010-8_7.

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

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Chongyang Ma, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2025a. doi: 10.52202/085713-4982. URL https://proceedings.neurips.cc//paper_files/paper/2025/hash/dae77d03bd51a5acfe8519848a3af6c9-Abstract-Datasets_and_Benchmarks_Track.html.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Yuan_Identity-Preserving_Text-to-Video_Generation_by_Frequency_Decomposition_CVPR_2025_paper.pdf.

Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, Tiankun Cao, Cheng Wang, Xiaotao Gu, Jie Tang, Dan Guo, and Meng Wang. Kaleido: Open-sourced multi-subject reference video generation model, 2025. URL https://arxiv.org/abs/2510.18573.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-id: Towards universal identity-preserving video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1927–1936, 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/HiGen/html/Zhong_Concat-ID_Towards_Universal_Identity-Preserving_Video_Synthesis_ICCVW_2025_paper.html.

A Analysis of Terminal Scoring and Selective Anchors

Setup and full resampling. Let ξ collect exogenous randomness, including conditioning data and sampling noise. Fix its distribution, the sampler, and an anchor distribution P for the analyzed update. The terminal output and population reward are

$$X_\theta(\xi) = x_N, \quad x_{k+1} = \Phi_k(x_k, \theta; \xi), \quad F(\theta) = \mathbb{E}_\xi R_\xi(X_\theta(\xi)), \tag{9}$$

where x_0 is independent of θ. Evaluating F after changing θ reruns the entire sampling chain; a cached terminal corresponds to the pre-update parameters. All update expectations below are conditional on the current training history, and θ denotes the current parameters.

Cache $\bar{x}_k = x_k(\theta; \xi)$ and define the leap $p_t(\vartheta; \bar{x}_t, \xi)$ with its anchor input held fixed. At $\vartheta = \theta$, write

$$A_t = \partial_\theta p_t, \qquad J = D_\theta X_\theta, \qquad \omega = \nabla_x R_\xi(X_\theta), \qquad u = \nabla F(\theta) = \mathbb{E}_\xi[J^\top \omega]. \tag{10}$$

The last identity assumes differentiation–expectation interchange. J includes shared-parameter effects through every sampling step, whereas A_t excludes derivatives through the cached anchor. Jacobians and reward gradients remain coupled inside expectations. Vector norms are Euclidean and matrix norms are induced operator norms.

Connector identities and proxy discrepancy. Let sg preserve forward values and have zero automatic derivative. At the current parameters, the connector satisfies

$$\begin{aligned} h_t &= p_t + \mathrm{sg}(X_\theta - p_t) = X_\theta, \\ D_\theta^{\mathrm{AD}} h_t &= A_t, \qquad \gamma_t = D_\theta^{\mathrm{AD}} R_\xi(h_t) = A_t^\top \omega, \\ \gamma_t - J^\top \omega &= (A_t - J)^\top \omega. \end{aligned} \tag{11}$$

Here reward differentiation is assumed faithful to R_ξ; detached scoring operations are treated below. These identities follow directly from the definition of stop-gradient and the chain rule, with the forward-value equality used only as a numerical relation.

Direct leap scoring gives $\gamma_t^{\mathrm{proxy}} = A_t^\top \nabla R_\xi(p_t)$. If the same reward has L_R-Lipschitz gradient on a neighborhood of the segment joining p_t and X_θ, then

$$\begin{aligned} \|\gamma_t^{\mathrm{proxy}} - \gamma_t\| &\le \|A_t\| L_R \|p_t - X_\theta\|, \\ \|\gamma_t^{\mathrm{proxy}} - J^\top \omega\| &\le \|(A_t - J)^\top \omega\| + \|A_t\| L_R \|p_t - X_\theta\|. \end{aligned} \tag{12}$$

Indeed, the direction difference is $A_t^\top[\nabla R_\xi(p_t) - \omega]$; apply the operator-norm and triangle inequalities. The connector removes the scoring-location term but retains Jacobian bias. The decomposition identifies the two error sources; since proxy-scoring and Jacobian errors can cancel, the ordering of actual total errors must be determined separately.

Endpoint decay of the raw reward gradient. Hold the cached $\bar{x}_t$ fixed and take $p_t = \bar{x}_t + c_t v_\theta(\bar{x}_t, t)$, where the scalar c_t is fixed conditional on the history and anchor t, and independent of θ. Define

$$B_t = \partial_\theta v_\theta(\bar{x}_t, t), \qquad A_t = c_t B_t, \qquad \gamma_t = c_t B_t^\top \omega. \tag{13}$$

Assume $\|B_t\| \le K_v$ and $\|\omega\| \le K_R$ almost surely, with finite constants uniform over the anchor family under consideration. Then, with expectations conditional also on t,

$$\begin{aligned} \|\gamma_t\| &\le |c_t| K_v K_R, \\ \mathbb{E}\|\gamma_t\|^2 &= c_t^2 \mathbb{E}\|B_t^\top \omega\|^2 \le c_t^2 K_v^2 K_R^2. \end{aligned} \tag{14}$$

Derivation. Differentiating with the cache and coefficient fixed gives Eq. (13). The operator-norm inequality gives the first bound; squaring and taking conditional expectation gives the second-moment statement. Thus $c_t \to 0$ under these uniform assumptions suppresses only the raw reward contribution, including its second moment.

Joint identity–motion window tradeoff. Fix a finite nonempty window $\mathcal{W}$ conditional on the current history, before drawing update randomness, and sample t uniformly from $\mathcal{W}$. For a fixed $\eta > 0$, let $\widehat{d}_t$ be the actual optimizer increment divided by η, including optimizer transformations and regularization. Let $z_\theta(\zeta)$ be a differentiable terminal motion descriptor from full resampling, with fixed-distribution evaluation randomness ζ conditionally independent of the anchor and update randomness. On descriptor space use the fixed weighted norm $\|z\|_{w,1} = w^\top |z|$, with $|z|$ taken elementwise and all entries of w positive. Writing $\mathbb{E}_t$ for expectation conditional also on t, define

$$\begin{aligned} a_t &= \langle \nabla F(\theta), \mathbb{E}_t \widehat{d}_t \rangle, \qquad M_t = \mathbb{E}_t \|\widehat{d}_t\|^2, \\ b_t &= \mathbb{E}_t \|D_\theta z_\theta(\zeta) \widehat{d}_t\|_{w,1}. \end{aligned} \tag{15}$$

Assume $M_t, b_t < \infty$ for every $t \in \mathcal{W}$. On a common neighborhood containing all realized update segments, assume F is L-smooth and the descriptor Taylor remainder obeys $\|z_{\theta+\eta'\widehat{d}_t}(\zeta) - z_\theta(\zeta) - \eta' D_\theta z_\theta(\zeta)\widehat{d}_t\|_{w,1} \le H\eta'^2 \|\widehat{d}_t\|^2/2$ for every $0 \le \eta' \le \eta$, almost surely, with common finite $L, H \ge 0$. Let $\bar{a}_\mathcal{W}$, $\bar{b}_\mathcal{W}$, and $\bar{M}_\mathcal{W}$ denote the averages of a_t, b_t, and M_t over $t \in \mathcal{W}$. Expectations below include the uniform anchor draw; the descriptor difference uses the same ζ before and after updating. Then

$$\begin{aligned} \mathbb{E}[F(\theta + \eta\widehat{d}_t)] - F(\theta) &\ge \eta \bar{a}_\mathcal{W} - \frac{L\eta^2}{2} \bar{M}_\mathcal{W}, \\ C_\mathcal{W} := \mathbb{E}\|z_{\theta+\eta\widehat{d}_t}(\zeta) - z_\theta(\zeta)\|_{w,1} &\\ \le \eta \bar{b}_\mathcal{W} + \frac{H\eta^2}{2} \bar{M}_\mathcal{W}. & \end{aligned} \tag{16}$$

Derivation. Conditional on t, smoothness gives the reward bound with a_t, M_t. Taylor expansion and the triangle inequality bound each descriptor difference by $\eta\|D_\theta z_\theta(\zeta)\widehat{d}_t\|_{w,1} + H\eta^2\|\widehat{d}_t\|^2/2$. Conditional expectation gives $\eta b_t + H\eta^2 M_t/2$; averaging both inequalities over uniform t gives Eq. (16).

For an identity reward F, these bounds express a finite-step tradeoff: favor high $\bar{a}_\mathcal{W}$, low $\bar{b}_\mathcal{W}$, and controlled $\bar{M}_\mathcal{W}$. Here $C_\mathcal{W}$ is expected per-sample motion-descriptor *disturbance*; the clean-prediction probe of §3.2, perturbed through the velocity, serves as an empirical proxy for these quantities.

Local improvement for the actual update. Consider $\theta^+ = \theta + \eta\widehat{d}$, with $\eta > 0$. Assume F is L-smooth, $L > 0$, on a common neighborhood containing every realized update segment, almost surely. Suppose

$$a = \langle u, \mathbb{E}\widehat{d} \rangle > 0, \qquad M = \mathbb{E}\|\widehat{d}\|^2 < \infty.$$

Then the full-resampling objective satisfies

$$\mathbb{E}[F(\theta^+)] - F(\theta) \ge \eta a - \frac{L\eta^2}{2} M. \tag{17}$$

Consequently, $0 < \eta < 2a/(LM)$ makes this lower bound positive, and the conservative choice $0 < \eta \le a/(LM)$ gives a lower bound of $\eta a/2$. These choices must also keep the update segments inside the smoothness neighborhood.

Derivation. Smoothness gives $F(\theta + \eta\widehat{d}) \ge F(\theta) + \eta\langle u, \widehat{d}\rangle - L\eta^2\|\widehat{d}\|^2/2$. Conditional expectation gives Eq. (17); substituting the step-size bounds gives both statements.

The direction $\widehat{d}$ is the actual optimizer increment divided by η. It therefore includes AdamW preconditioning and momentum, teacher-velocity regularization, clipping, and decoupled weight decay when present. Alignment must hold for this combined direction. An increase in the regularized objective arises from its reward and penalty components jointly, so the reward component must be accounted for separately. If the direction itself depends on the chosen step size, a and M must be evaluated for that choice. An increment already containing the learning rate is used as is.

Reward SGD. Let $\widehat{d} = \bar{\gamma} = n^{-1}\sum_{i=1}^n \gamma_i$, $\mu = \mathbb{E}\bar{\gamma}$, and suppose

$$\|\mu - \kappa u\| \le \delta\|u\|, \quad \kappa > \delta \ge 0, \quad \mathbb{E}\|\bar{\gamma} - \mu\|^2 \le V/n.$$

Under the same smoothness assumptions,

$$\mathbb{E}[F(\theta^+)] - F(\theta) \geq \eta(\kappa - \delta)\|u\|^2 - \frac{L\eta^2}{2}\left[(\kappa + \delta)^2\|u\|^2 + \frac{V}{n}\right]. \tag{18}$$

Derivation. Cauchy–Schwarz yields $\langle u, \mu\rangle \geq (\kappa - \delta)\|u\|^2$ and $\|\mu\| \leq (\kappa + \delta)\|u\|$. The identity $\mathbb{E}\|\bar{\gamma}\|^2 = \|\mu\|^2 + \mathbb{E}\|\bar{\gamma} - \mu\|^2$ completes the derivation. For $u \neq 0$, set $a_0 = (\kappa - \delta)\|u\|^2$ and $M_0 = (\kappa + \delta)^2\|u\|^2 + V/n$. Any $0 < \eta < 2a_0/(LM_0)$ gives a positive bound; $\eta \leq a_0/(LM_0)$ gives at least $\eta a_0/2$. No strict improvement follows at $u = 0$.

Independent versus shared anchors. Assume $t \sim P$ is independent of ξ. Define $\mu_t = \mathbb{E}_\xi \gamma_{\xi,t}$, $V_t = \mathbb{E}_\xi\|\gamma_{\xi,t} - \mu_t\|^2$, and $\mu = \mathbb{E}_t\mu_t$. For n independent sample–anchor pairs, and for independent samples sharing one random batch anchor, respectively, total variance gives

$$\begin{aligned} \mathbb{E}\|\bar{\gamma}_{\text{ind}} - \mu\|^2 &= \frac{\mathbb{E}_t V_t + \mathbb{E}_t\|\mu_t - \mu\|^2}{n}, \\ \mathbb{E}\|\bar{\gamma}_{\text{shared}} - \mu\|^2 &= \frac{\mathbb{E}_t V_t}{n} + \mathbb{E}_t\|\mu_t - \mu\|^2. \end{aligned} \tag{19}$$

Thus shared-anchor variability need not vanish with batch size. The SGD variance assumption must match the actual batching scheme; the local-improvement bound, Eq. (17), instead uses its actual second moment directly.

Detached scoring and hard minima. Input derivatives persist even with frozen network parameters. If detection coordinates or output-dependent soft weights are detached, let ω_{AD} be the resulting input derivative and $\gamma_t^{\text{AD}} = A_t^\top \omega_{\text{AD}}$. Wherever the mathematical reward is differentiable,

$$\gamma_t^{\text{AD}} - \kappa J^\top \omega = (A_t - \kappa J)^\top \omega + A_t^\top(\omega_{\text{AD}} - \nabla R_\xi(X_\theta)). \tag{20}$$

Both terms enter the alignment budget. For example, their expected norm sum being at most $\delta\|u\|$ suffices for the reward-SGD bound with the implemented mean direction. The gap between a fixed-detection surrogate and the reward obtained by redetecting the resampled output is captured by these two norms.

A hard minimum of smooth member scores is locally smooth when its minimizing member stays unique and the relevant branches have uniform smoothness bounds. At ties, differentiability requires agreement of all active gradients; this alone does not establish a neighborhood-wide Lipschitz gradient. For $\min(y, -y) = -|y|$ at zero, following either active gradient decreases the minimum. Stable detection, matching, valid-frame sets, nondegenerate normalizations, and suitable derivative bounds can justify local analysis; almost-everywhere differentiability alone cannot. Alternatively, F may be independently shown smooth. Replacing hard-min by softmin defines a different objective and does not smooth underlying detection branches.

Conditional geometry of soft reference aggregation. Fix a nonempty frame set $\mathcal{F}$, with verifier index $m \in \{\text{A}, \text{B}, \text{C}\}$ and reference index j. Let generated embeddings $\hat{g}_{fm}$ and reference embeddings $\hat{r}_{jm}$ be unit vectors, and let $\alpha_{fmj} \geq 0$, $\sum_j \alpha_{fmj} = 1$. The following are two conditional reward definitions used for the analysis, stated independently of any particular implementation.

For a *normalized soft prototype*, define

$$\bar{r}_{fm} = \frac{\tilde{r}_{fm}}{\|\tilde{r}_{fm}\|}, \quad \tilde{r}_{fm} = \sum_j \alpha_{fmj}\hat{r}_{jm}, \qquad s_{fm}^{\text{proto}} = \hat{g}_{fm}^\top \bar{r}_{fm}, \qquad \chi_{fm}^{\text{proto}} = \|\hat{g}_{fm} - \bar{r}_{fm}\|^2,$$

with unnormalized prototype norms $\|\tilde{r}_{fm}\|$ bounded away from zero on the analyzed neighborhood. For a *weighted pairwise score*, define instead

$$s_{fm}^{\text{pair}} = \sum_j \alpha_{fmj}\hat{g}_{fm}^\top \hat{r}_{jm}, \qquad \chi_{fm}^{\text{pair}} = \sum_j \alpha_{fmj}\|\hat{g}_{fm} - \hat{r}_{jm}\|^2.$$

For either choice $\ell \in \{\text{proto}, \text{pair}\}$, the unit-vector identity gives $s_{fm}^\ell = 1 - \chi_{fm}^\ell/2$, hence

$$R_\ell = \frac{1}{|\mathcal{F}|}\sum_{f \in \mathcal{F}} \min_m s_{fm}^\ell = 1 - \frac{\bar{\chi}_\ell}{2}, \qquad \bar{\chi}_\ell = \frac{1}{|\mathcal{F}|}\sum_{f \in \mathcal{F}} \max_m \chi_{fm}^\ell. \tag{21}$$

If the analyzed reward equals R_ℓ, any lower bound on the expected reward gain implies a decrease of at least twice that bound in the corresponding expected $\bar{\chi}_\ell$. The two distances must be handled separately. Output-dependent weights remain part of the reward derivative unless their omission is accounted for by Eq. (20).

Example (correct scoring without alignment). Let $t \in [0, 1]$ denote integration time, $x_0 = 0$, and $v_\theta(t) = (1 - 8t)\theta$. Two Euler steps of size $1/2$, evaluated at $t = 0$ and $t = 1/2$, give

$$x_1 = \tfrac{1}{2} v_\theta(0) = \tfrac{1}{2}\theta, \qquad X_\theta = x_1 + \tfrac{1}{2} v_\theta(1/2) = \tfrac{1}{2}\theta - \tfrac{3}{2}\theta = -\theta.$$

The first-anchor leap over the full remaining time is $p_\theta = x_0 + v_\theta(0) = \theta$. Choose $R(x) = \sin x = \langle(\cos x, \sin x), (0, 1)\rangle$, a unit-feature similarity. At $\theta = 0$, $p = X = 0$, $\omega = 1$, $A = 1$, and $J = -1$. The connector therefore gives $\gamma = 1$, while $F(\theta) = -\sin\theta$ has $u = -1$. Updating to $\theta^+ = \eta$ yields

$$F(\theta^+) - F(0) = -\sin\eta < 0, \qquad 0 < \eta < \pi.$$

The scoring-location error is zero. Opposing shared-parameter Jacobians alone cause failure, so terminal-value consistency alone does not replace the alignment condition.

B Training and Inference Configuration

Optimization. Main experiments run on 128 H20 GPUs with context parallel 2; the global batch is 64. Each batch keeps a fixed ratio of single- and multi-person training clips. The DiT is fully fine-tuned in bf16 with 8-bit AdamW (Dettmers et al., 2022) (learning rate 2×10^{-5}, cosine decay to 2×10^{-6}, 10 warmup steps); the frozen teacher shares the same architecture and provides the teacher-velocity target.

Resolution and sampling. Training videos are 480p/81-frame, bucketed by aspect ratio (landscape 480×864, portrait 864×480); evaluation videos are 480p/121-frame. HunyuanVideo-1.5 is the CFG-free variant; Wan2.2 uses its standard CFG (Ho & Salimans, 2022) dual-pass.

Objective and windows. The reward weight is λ=0.1; the velocity-penalty weight β is 10 (HunyuanVideo-1.5 and MiniMax-H3), 5 (Wan2.2/VACE), and 3 (12-step). The STAR windows are $[36, 47]$ (HunyuanVideo-1.5), $[39, 47]$ (Wan2.2), $[37, 47]$ (VACE), $[8, 10]$ (12-step), and $[38, 44]$ (MiniMax-H3). MiniMax-H3 instead uses bf16 AdamW with stochastic rounding (Adam betas $(0.9, 0.95)$) and a smaller learning rate of 5×10^{-6}; all other backbones use the settings above. At every iteration each training sample draws its own anchor uniformly within the window (one per context-parallel group), i.e., the independent-anchor case of Eq. (19); the identity-forward connector routes the gradient only through the anchor leap.

Reward computation. The connector is decoded at the true terminal value. InsightFace `buffalo_l` (InsightFace Contributors, 2026) supplies face detections; each member embeds the same valid subject–frame crops. Soft reference prototypes are normalized before scoring, then the three-member minimum is taken per frame, followed by frame and subject means, as in Eq. (1).

Regularization, dynamics, and calibration. The teacher-velocity penalty is the mean squared difference between policy and teacher velocities; when both transitions are Gaussian with a shared isotropic variance and means that differ by a step-dependent multiple of the velocity difference, it equals their KL divergence up to a step-dependent factor. The dynamic descriptor is the per-subject temporal standard deviation of yaw, pitch, and the 212-dimensional expression descriptor (averaged over coordinates), averaged over subjects; the weights w_p, w_e in Eq. (4) equalize the RMS scales of the three terms over all probe records and stay fixed across timesteps. The windows above are calibrated once per backbone from the probe curves and kept fixed during training; the released reference implementation also provides a threshold-based recalibration that returns the longest contiguous interval satisfying $q(k) \leq 45$ and the identity and disturbance thresholds.

C Verifier Training Details

We fine-tune the later layers of the public ArcFace-R100 on 2.9k real identities (≈72k face crops), with ≈55k generated face crops as hard negatives; test identities are disjoint from training identities. The loss combines a pose/expression-aware supervised contrastive term, distillation toward the

frozen R100, a margin on the real–generated similarity gap, and generated identity-drift frames as per-identity negatives. The three members differ in the breadth of the negative pool and the strength of the drift term; the reward is $\min(s_{\mathrm{A}}, s_{\mathrm{B}}, s_{\mathrm{C}})$ with soft aggregation over the face pool.

D HQ4000 Construction Pipeline

1. *Retrieval and detection*: hundreds of thousands of candidate clips are retrieved from an internal video library; frames are sampled and faces detected with InsightFace `buffalo_l` (InsightFace Contributors, 2026). Each kept clip is $\approx$5 s long and contains one or more clearly visible subjects.
2. *Filtering*: a VLM judges video quality and occlusion, and only clips with sufficient expression and angle variation are kept—every retained clip has a face-angle variation above 70°, so that post-training lands in the large-pose, large-expression regime.
3. *Per-face annotation*: for each clip, 25 pool frames are sampled over the angle and expression range. Yaw/pitch are annotated with the 3D-68 pose model (`1k3d68`) (InsightFace Contributors, 2026); 212-dimensional expression descriptors and ArcFace (Deng et al., 2019) 512-dimensional embeddings are stored alongside. The expression descriptor flattens the 106 `2d106det` landmarks after centering them on the face box and dividing by its longer side. The reference and generated faces use the same detector and coordinate ordering.
4. *Multi-person handling*: for multi-person clips (2–7 people), a separate pool is built per subject under `subjects/person_*/`, and at reward time each detected face is matched to its subject's pool by identity, so that every subject is supervised with its own references. Multi-subject supervision is the setting this dataset adds over single-subject ones.
5. *Caption*: Qwen3.5-397B-A17B (Qwen Team, 2026) writes a formatted caption per clip as the I2V prompt, so that the generated video matches the original clip's facial motion and the faces land within the pool's coverage.

Each sample stores `ref.png`, the source clip, the caption, and per-subject `face_pool.npz` with fields `embeddings` ($n_{\mathrm{pool}}\times512$), `yaws`/`pitches` (degrees), and detection scores, together with the $n_{\mathrm{pool}}\times212$ expression descriptors of the pool. In total the dataset covers **5,115** identity-supervised subjects and **125k+** face crops ($\approx$24 per subject). The training split has 4,000 clips and the test split is packaged as test500 (§5.1); test identities are disjoint from training. The same pipeline can be applied to similar video sources to reconstruct a comparable dataset.

E Single- vs. Multi-Person Breakdown

Test500 contains 357 single-person and 143 multi-person clips (2–4 people per clip). Table 6 reports both subsets: STAR improves identity on both.

Table 6: Single-/multi-person breakdown on test500.

	min3 FaceSim		raw ArcFace	
Arm	single (357)	multi (143)	single (357)	multi (143)
HY1.5 (50-step) base	0.5934	0.5441	0.4728	0.4172
+ STAR	**0.7519**	**0.6575**	**0.6207**	**0.5326**
HY1.5 (12-step) base	0.5823	0.5288	0.4982	0.4391
+ STAR	**0.7959**	**0.7022**	**0.6751**	**0.5856**
Wan2.2 (50-step) base	0.5522	0.5022	0.3860	0.3318
+ STAR	**0.7473**	**0.6578**	**0.5949**	**0.5075**
MiniMax-H3 base	0.6744	0.6048	0.5101	0.4531
+ STAR	**0.7417**	**0.6599**	**0.5885**	**0.5196**

Table 7: Mechanisms and adaptation requirements of identity reward post-training methods.

Method	Reward interface	Gradient mechanism	Adaptation considerations
IPRO	RGB ArcFace	Continuous K-step suffix	Longer gradient horizon increases retained/recomputed work
Identity-GRPO	Preference-trained VLM	Group-relative policy update	Reward calibration on the target output distribution
DreamVideo-Omni	Learned latent identity reward	Intermediate-step reward	Compatibility with target VAE coordinates
STAR	Adapted RGB verifier	One anchor in a calibrated window	Verifier, reference pools, and backbone-specific calibration

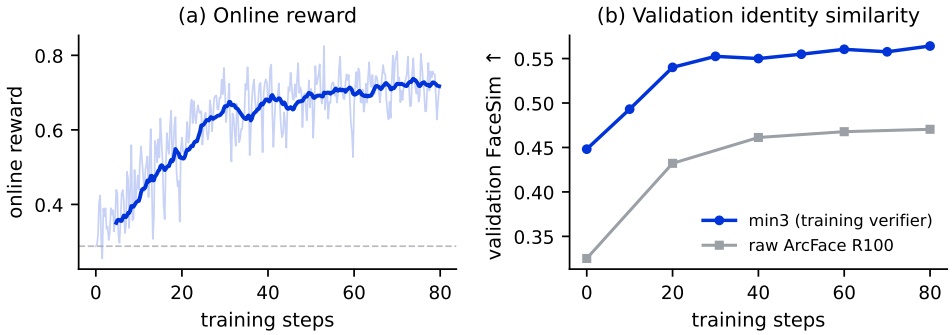


Figure 6: Training curves. (a) Online min3 reward (light: raw; solid: 5-step average; dashed: initial). (b) Validation FaceSim on 32 held-out prompts (min3 and raw ArcFace R100).

F ADDITIONAL COMPARISONS AND TRAINING BEHAVIOR

F.1 METHOD COMPARISON

Table 7 above compares reward interfaces, gradient mechanisms, and adaptation needs.

F.2 TRAINING CURVES

Both curves in Figure 6 rise smoothly and plateau without oscillation or collapse, and the validation gain also holds under raw ArcFace R100.

F.3 REFL-STYLE CONNECTOR ABLATION

We compare STAR with a ReFL-style variant (Xu et al., 2023) that scores the one-step prediction p_t directly instead of $h_t = p_t + \mathrm{sg}(x_N - p_t)$, keeping the window $[36, 47]$, $\beta = 10$, min3 reward, data, and optimizer fixed. Table 8 shows comparable FaceSim (0.7271 vs. 0.7249) but lower dynamic retention without the connector (0.85 vs. 0.89): scoring the true terminal keeps identity while preserving more motion. The connector identities in Appendix A separate this scoring-location effect from the remaining Jacobian approximation.

Table 8: ReFL-style connector ablation (HunyuanVideo-1.5, test500; FaceSim is min3).

Method	FaceSim ↑	yaw_std	pitch_std	expr_std	Dyn× ↑
base	0.5793	16.86	11.84	0.0479	1.00
ReFL (no connector)	0.7271	14.50	9.94	0.0405	0.85
STAR	0.7249	15.10	10.37	0.0425	**0.89**