PISCES: Annotation-free Text-to-Video Post-Training via Optimal Transport-Aligned Rewards

Minh-Quan Le [*1,2] **Gaurav Mittal** [*1] **Cheng Zhao** [1] **Xianfeng David Gu** [2] **Dimitris Samaras** [2] **Mei Chen** [†1]

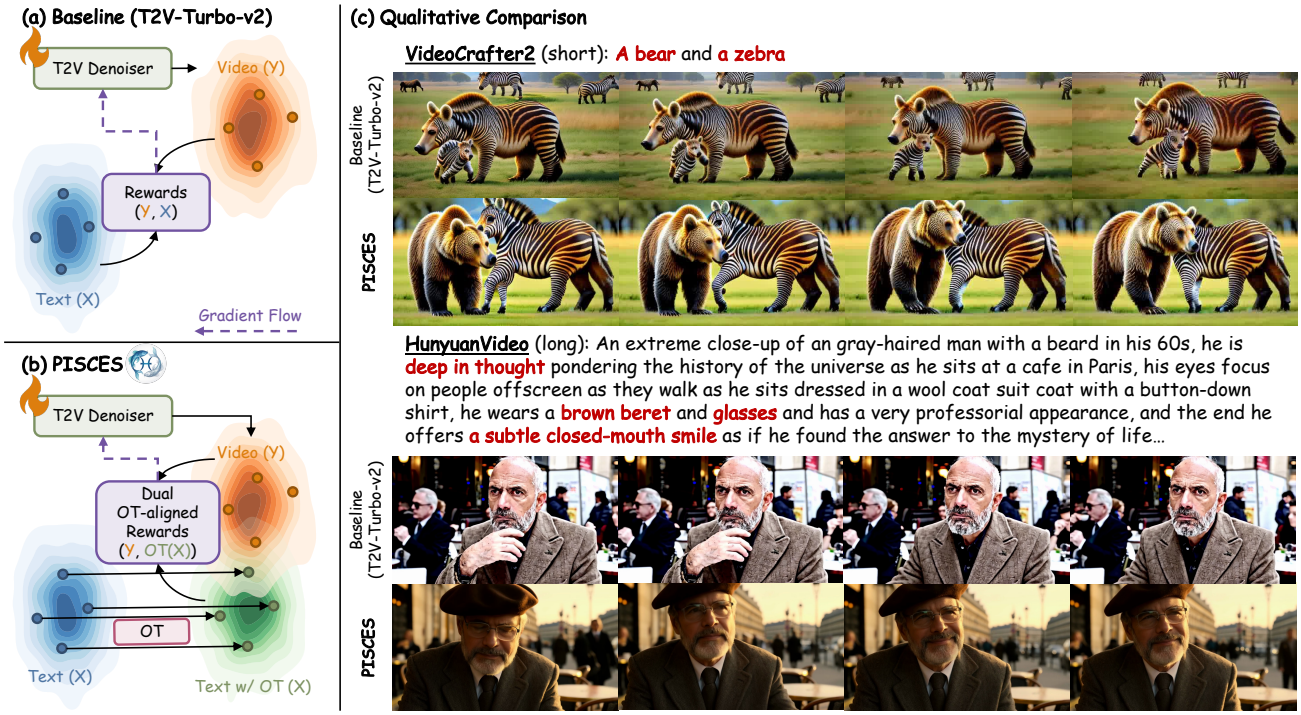

Figure 1. (a) Baseline (T2V-Turbo-v2) defines rewards over pre-trained VLM text-video embeddings, which suffer from distributional misalignment. (b) PISCES T2V post-training addresses this by formulating reward supervision over an OT-aligned embedding space. We propose a novel Dual OT-aligned Rewards module that aligns text embeddings to the video space, enabling effective visual and semantic alignment. (c) Compared to the baseline, PISCES improves visual quality (temporal coherence, photorealism) and semantic fidelity (object count, attributes) on both short-video (VideoCrafter2) and long-video (HunyuanVideo) generation.

Abstract

Text-to-video (T2V) generation aims to synthesize videos with high visual quality and temporal consistency that are semantically aligned with input text. Reward-based post-training has emerged as a promising direction to improve the quality and semantic alignment of generated videos. However, recent methods either rely on large-scale human preference annotations or operate on misaligned embeddings from pre-trained vision-language models, leading to limited scalability or suboptimal supervision. We present PISCES, an annotation-free post-training algorithm that addresses these limitations via a novel Dual Optimal Transport (OT)-aligned Rewards module. To align reward signals with human judgment, PISCES uses OT to bridge text and video embeddings at both distributional and discrete token levels, enabling reward supervision to fulfill two objectives: (i) a Distributional OT-aligned Quality Reward that captures overall visual quality and temporal coherence; and (ii) a Discrete Token-level OT-aligned Semantic Reward that enforces semantic, spatio-temporal correspondence between text and video tokens. To our knowledge, PISCES is the first to improve annotation-free reward supervision in generative post-training through the lens of OT. Experiments on both short- and long-video generation show that PISCES outperforms both annotation-based and annotation-free methods on

[*]Equal contribution. Done during MQ's internship at Microsoft. [†]Currently at Dolby Labs. [1]Microsoft [2]Stony Brook University .

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

VBench across Quality and Semantic scores, with human preference studies further validating its effectiveness. We show that the Dual OT-aligned Rewards module is compatible with multiple optimization paradigms, including direct backpropagation and reinforcement learning fine-tuning. Project page: https://roar-ai.github.io/pisces

1. Introduction

Text-to-video (T2V) generation (Kong et al., 2025; RunwayML, 2024) aims to synthesize videos from textual descriptions such that they appear realistic, temporally consistent, and accurately reflect the prompt. T2V has broad applications in multimedia content creation, robotics, and accessibility. While T2V performance is inherently subjective and judged by human preferences, recent benchmark (Huang et al., 2024) formalizes evaluation along two main dimensions – *Quality score*, for the visual quality and temporal consistency; and *Semantic score*, factoring the correspondence of generated videos to text prompts.

Rapid advances in diffusion and flow matching models (Podell et al., 2024; Esser et al., 2024) and Vision-Language Models (VLMs) (Chung et al., 2023; Sun et al., 2024; GLM et al., 2024) have enabled the development of recent T2V models (Pika Labs, 2023; RunwayML, 2024; Chen et al., 2024; Kong et al., 2025). To further improve existing T2V models (Chen et al., 2024; Kong et al., 2025), particularly in terms of video-text misalignment in the denoisers, reward-based post-training (Li et al., 2025a; Liu et al., 2025b) has been introduced that provides additional supervision via specially designed rewards to the denoiser.

Reward-based T2V post-training methods can be either Annotation-based or Annotation-free. Annotation-based approaches (Liu et al., 2025b; Yang et al., 2026; Wang et al., 2025b) collect large-scale human preference datasets, where annotators judge generated video pairs on quality and semantics, which are used to train a reward model or via Direct Preference Optimization (DPO) (Rafailov et al., 2023; Wallace et al., 2024) for post-training. Although effective and serving as existing SoTA, these annotation-based methods cannot easily scale because they need high-quality preference-based annotations. Another line of work explores Annotation-free rewards, where supervision is derived from pre-trained VLMs rather than human labels (Li et al., 2024; 2025a). While these approaches do not need large-scale human annotations, their performance is not on par with the Annotation-based techniques. We aim to achieve the best of both worlds by asking: *Can an annotation-free T2V post-training method match, or even outperform, annotation-based approaches*?

From a review of annotation-free approaches, we identify reliance on pre-trained vision–language models (VLMs) for reward supervision as a key limitation. VLMs are trained with non-distributional objectives, such as pointwise matching (Chen et al., 2020) and contrastive learning (Radford et al., 2021), that may not adequately align text with the real-video distribution, consistent with the patterns in Table 4 and Figure 6. This results in both quality and semantic issues, as shown in Figure 1c, such as failure to ensure the correct number and attributes of objects (*e.g.*, "a zebra and a bear", "wearing a brown beret and glasses") or failing to capture motion descriptors (*e.g.*, "closed-mouth smile").

We posit that, for annotation-free reward supervision to mimic human preferences, the real-video distribution must be better aligned with the text distribution, which represents the space of human instructions/preferences, without altering the video distribution's semantic structure, and the derived rewards should reflect human judgments of text-to-video outputs on the dual of quality and semantics. We introduce PISCES[1], an annotation-free T2V post-training algorithm that includes a novel **Dual Optimal Transport-aligned Rewards** module (Figure 1b). Leveraging Optimal Transport (OT) (Villani, 2009; Cuturi, 2013), we tailor PISCES specifically for T2V post-training by enhancing text-video alignment at both the distribution and the token level to simultaneously improve both the visual quality and semantic consistency. For this, the Dual Rewards module comprises: (i) a **Distributional OT-aligned Quality Reward**, which learns a distributional OT map to transform text embeddings into the real-video embedding space while preserving their internal structure and enforcing temporal consistency and visual quality; and (ii) a **Discrete Token-level OT-aligned Semantic Reward**, which constructs a semantic spatio-temporal cost matrix over text and video tokens and solves a partial OT problem with an entropic Sinkhorn solver (Cuturi, 2013), to supervise correspondence by aligning text tokens with the most semantically, spatially, and temporally consistent video regions.

We validate PISCES on both short-video (VideoCrafter2 (Chen et al., 2024)) and long-video (HunyuanVideo (Kong et al., 2025)) generator (Figure 1c) via VBench (Huang et al., 2024) as well as human evaluation. We show that our Dual OT-aligned Rewards module is applicable across different optimization paradigms, including direct backpropagation (gradient backpropagation through reward models) and reinforcement learning (RL) fine-tuning (GRPO (DeepSeek-AI et al., 2026; Liu et al., 2025a)). In doing so, we find that PISCES can significantly outperform all existing reward-based post-training approaches (both Annotation-based and Annotation-free). Through careful realignment of the text-video space, PISCES demonstrates that an annotation-free T2V

[1]In astrology, Pisces is symbolized by two fish, signifying balance across realms. PISCES echoes this by aligning text and video through complementary quality and semantic rewards.

post-training method can outperform Annotation-based approaches, making it a much stronger alternative at scale. Our key contributions are:

- We introduce PISCES, a novel annotation-free post-training framework for T2V generation. For the first time, we identify a core bottleneck in existing VLM-based rewards operating on misaligned text-video embeddings and address this by leveraging OT to align embeddings, enabling reward supervision in a semantically meaningful, structure-preserving space.
- PISCES defines a novel Dual OT-aligned Rewards module comprising: (1) a *Distributional OT-aligned Quality Reward*, capturing overall visual quality and temporal consistency; and (2) a *Discrete Token-level OT-aligned Semantic Reward*, targeting localized text-video alignment for semantic consistency.
- PISCES outperforms both annotation-based and annotation-free T2V post-training methods on both Semantic and Quality Scores for short and long video generation, as validated by automatic metrics and human evaluations. We show that OT-aligned rewards are applicable to multiple optimization strategies.

2. Related Work

Reward-based Post-Training for T2V. In the image domain, reward models such as HPSv3 (Ma et al., 2025), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and Hummingbird (Le et al., 2025a) have proven effective for aligning generations with text prompts. Extending to video, annotation-based approaches train on large-scale human preference datasets, including VideoReward (Liu et al., 2025b), Dual-IPO (Yang et al., 2026), and UnifiedReward (Wang et al., 2025b). While effective, these methods incur high annotation costs and suffer from limited scalability. Orthogonally, annotation-free methods leverage verifiable rewards (Le et al., 2025b) or pre-trained VLMs such as ViCLIP (Wang et al., 2024) and InternVideo2 (Wang et al., 2025a), with cosine-similarity rewards adopted in T2V-Turbo (Li et al., 2024), T2V-Turbo-v2 (Li et al., 2025a), while InstructVideo (Yuan et al., 2024) still relies on image-text rewards. However, these rewards operate on misaligned text–video embedding spaces, reducing their effectiveness to perform on par with annotation-based methods. We identify this reward misalignment as the core bottleneck in T2V post-training and propose PISCES, the first to explore aligning embeddings via OT in generative post-training. Our framework introduces dual distributional and token-level OT-aligned rewards, enabling scalable and effective annotation-free supervision.

Optimal Transport. Optimal Transport (OT) provides a principled framework for aligning probability distributions and has been widely applied in machine learning tasks such as domain adaptation (Katageri et al., 2024), generative modeling (Tong et al., 2024; Li et al., 2023), and cross-modal retrieval (Han et al., 2024; Izquierdo & Civera, 2024). Neural Optimal Transport (NOT) (Korotin et al., 2023) further offers a scalable alternative by learning explicit transport maps via neural networks. Recent works have also leveraged discrete OT for alignment in vision tasks: (Xie et al., 2025) formulate disentangled OT for visual–concept relations, while (Liu et al., 2025c) integrates OT into query reformation for temporal action localization. Other efforts include HOTS3D (Li et al., 2025b), which aligns text and image features using spherical OT, and OT-CLIP (Shi et al., 2024), which reframes CLIP training and inference as OT problems. Despite these advances, prior work has not explored OT in the context of reward modeling for generative T2V post-training. PISCES is the first to address misaligned pre-trained VLM text-video embeddings for annotation-free T2V rewards, and introduces a Dual OT-aligned Rewards module capturing both distributional alignment and discrete token-level text-video correspondence.

3. Method

Fig 2 provides an overview of PISCES's Dual OT-aligned Rewards module and T2V post-training algorithm.

3.1. Distributional OT-aligned Quality Reward

When humans evaluate the *quality* of a generated video, they attend to global properties such as realism, motion coherence, and overall visual consistency – essentially asking whether the video could plausibly belong to the distribution of real-videos. To mimic this process in an annotation-free manner, we align text embeddings onto the manifold of real-video embeddings before defining the reward.

For this, we formulate the distributional alignment as a Monge–Kantorovich OT problem. OT (Villani, 2009) provides a principled framework for aligning two probability distributions $\mu \in \mathcal{P}(\mathcal{Y})$ and $\nu \in \mathcal{P}(\mathcal{X})$ by finding a transport map $\mathbf{T} : \mathcal{Y} \to \mathcal{X}$ that pushes μ to ν (i.e., $\mathbf{T}_\sharp \mu = \nu$) while minimizing a transport cost $\mathbf{c}(\mathbf{y}, \mathbf{x})$. Given text embeddings $\mathcal{Y}$ and real-video embeddings $\mathcal{X}$ extracted from a pre-trained VLM, we train an OT map $\mathbf{T} : \mathcal{Y} \to \mathcal{X}$ using NOT (Korotin et al., 2023) with the objective:

$$\sup_{f} \inf_{\mathbf{T}} \int_{\mathcal{X}} f(\mathbf{x}) d\nu(\mathbf{x}) + \int_{\mathcal{Y}} \left(\mathbf{c}(\mathbf{y}, \mathbf{T}(\mathbf{y})) - f(\mathbf{T}(\mathbf{y}))\right) d\mu(\mathbf{y}), \tag{1}$$

where $\mathbf{c}(\mathbf{y}, \mathbf{x}) = \|\mathbf{y} - \mathbf{x}\|^2$ is the transport cost. We implement this via iterative optimization of the transport map $\mathbf{T}_\psi$ and potential function f_ω parameterized by neural networks, as shown in Algorithm 2 in Appendix B. The resulting OT-aligned embeddings $\mathbf{T}^\star(\mathbf{y})$ reduce distributional mismatch while preserving the semantic structure of the embedding space (see Table 4 and Figure 6).

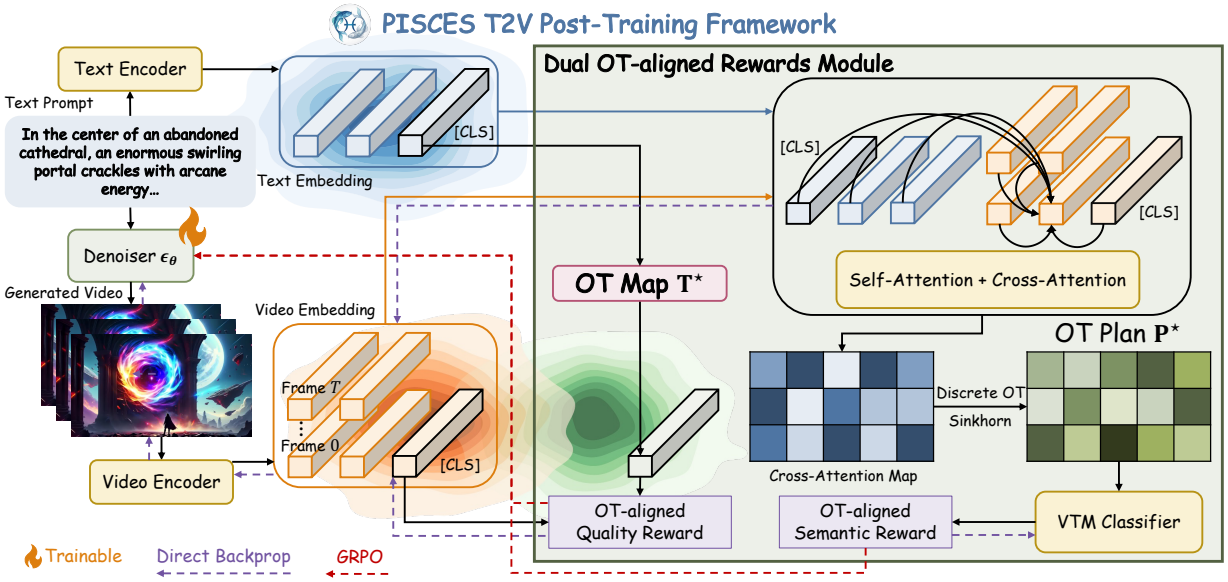

Figure 2. PISCES **T2V Post-Training.** We introduce a Dual OT-aligned Rewards module: (i) a distributional OT map $\mathbf{T}^\star$ for Quality Reward via [CLS] representation similarity, and (ii) a discrete OT plan $\mathbf{P}^\star$ with spatio-temporal constraints for Semantic Reward via a Video-Text Matching (VTM) classifier. The rewards module provides supervision for fine-tuning the T2V denoiser and is applicable with direct backpropagation and RL fine-tuning (GRPO).

Once the text distribution is aligned with the real-video distribution, comparing an OT-aligned text embedding $\mathbf{T}^\star(\mathbf{y})$ with a generated video embedding $\hat{\mathbf{x}}$ becomes equivalent to comparing a real-video embedding $\mathbf{x}^{\text{real}}$ with $\hat{\mathbf{x}}$. The OT map thus projects text embeddings into the real-video manifold, making $\mathbf{T}^\star(\mathbf{y})$ a proxy for $\mathbf{x}^{\text{real}}$. With this intuition, we define the Quality Reward (Fig 2) as the cosine similarity between global representation [CLS] tokens:

$$\mathcal{R}_{\text{OT-quality}} = \frac{\mathbf{T}^\star(\mathbf{y}_{[\text{CLS}]})^T \cdot \hat{\mathbf{x}}_{[\text{CLS}]}}{\|\mathbf{T}^\star(\mathbf{y}_{[\text{CLS}]})\|\|\hat{\mathbf{x}}_{[\text{CLS}]}\|} \approx \frac{(\mathbf{x}^{\text{real}}_{[\text{CLS}]})^T \cdot \hat{\mathbf{x}}_{[\text{CLS}]}}{\|\mathbf{x}^{\text{real}}_{[\text{CLS}]}\|\|\hat{\mathbf{x}}_{[\text{CLS}]}\|}. \tag{2}$$

Cosine similarity provides a natural coherence signal by comparing the *direction* of embeddings, making it robust to scale or style differences while capturing structural consistency. After OT projects text embeddings into the real-video manifold, cosine becomes a geometry-respecting measure of quality, evaluating whether generated videos point in the same "quality direction" as real ones.

3.2. Discrete Token-level OT-aligned Semantic Reward

When judging semantic fidelity in T2V, humans implicitly ask whether the prompt's keywords are actually reflected in the generated video. To mimic this in T2V post-training, we introduce a token-level reward based on Partial OT (POT).

To facilitate a strong semantic alignment, we integrate discrete POT into the cross-attention layers of pre-trained VLM InternVideo2 (Wang et al., 2025a). Vanilla cross-attention, however, often fails to capture precise multimodal correspondences: it operates directly on misaligned embeddings and distributes attention diffusely across irrelevant patches, as seen in Fig 5. Without a mechanism to enforce selective, structured grounding, important tokens may fail to connect to the right visual regions. To address this problem, we design a novel mechanism which augments attention with a POT-guided transport plan that enforces semantic, temporal, and spatial consistency between text and video tokens. For each cross-attention head, we construct a cost matrix between text tokens $\mathbf{y}$ and video patch tokens $\hat{\mathbf{x}}$ comprising three components specifically designed for T2V rewards:

Semantic similarity: $1 - \cos(\mathbf{y}_i, \hat{\mathbf{x}}_j)$, encouraging tokens with similar meaning to align.

Temporal constraint: $|\tau(\mathbf{y}_i) - t_j|$, where $\tau(\mathbf{y}_i) = \sum_k \mathbf{A}_{ik} * t_k$ is the expected frame index of text token i under attention $\mathbf{A}$, and t_j is frame index of video patch j.

Spatial constraint: $|\pi(\mathbf{y}_i) - s_j|_2$, where $\pi(\mathbf{y}_i) = \sum_k \mathbf{A}_{ik} * s_k$ is the expected 2D position of text token i (under attention $\mathbf{A}$) and s_j is the spatial coordinate of video patch j on the frame grid.

The final cost matrix is: $\mathbf{C}_{ij} = \text{semantic}(i,j) + \gamma \cdot \text{temporal}(i,j) + \eta \cdot \text{spatial}(i,j)$, with γ, η balancing temporal and spatial penalties. We then solve a partial OT problem on this cost matrix $\mathbf{C}$ via an entropic Sinkhorn solver (Cuturi, 2013) with fraction-of-mass $m = 0.9$, as shown in Algorithm 3. This produces a transport plan $\mathbf{P}^\star$ that softly matches each text token to a subset of video tokens, rather than forcing full mass transport. To integrate this into InternVideo2, we propose to inject $\mathbf{P}^\star$ into the vanilla attention $\mathbf{A}$ via log-space fusion, a lightweight, differentiable mechanism. This yields updated cross-attention probabilities $\tilde{\mathbf{A}}$ that combines standard attention with POT-guided structure. Formally, the updated cross-attention map is:

$$\tilde{\mathbf{A}} \propto \exp\big(\log(\mathbf{A} + \varepsilon) + \log(\mathbf{P}^\star + \varepsilon)\big). \tag{3}$$

This fusion preserves differentiability through $\mathbf{A}$ while treating $\mathbf{P}^\star$ as a structural prior. Finally, the POT-refined

features are passed into the pre-trained Video-Text Matching (VTM) classifier of InternVideo2, which outputs two logits for positive and negative matches. The positive logit after softmax (idx $= 1$) provides the Semantic Reward:

$$\mathcal{R}_{\text{OT-semantic}} = \texttt{softmax}\left(\texttt{VTM}\left[\tilde{\mathbf{A}} \cdot \hat{\mathbf{x}}\right]\right)_{\text{idx}=1}. \quad (4)$$

Our discrete POT-based Semantic Reward captures human selectivity: not every word needs to be grounded, and important tokens are matched to relevant patches. The spatio-temporal cost further constrains *where* and *when* content should appear. Together, this provides a reward signal evaluating text–video correspondence in a human-aligned way.

3.3. Post-Training

Direct Backpropagation. We integrate the Dual OT-aligned Rewards module into consistency distillation (Song et al., 2023; Wang et al., 2023b; Lu & Song, 2025) for efficient refinement, optimizing the denoiser:

$$\mathcal{L}_{\text{direct}} = \mathcal{L}_{\text{CD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \phi) - \mathcal{R}_{\text{OT-quality}} - \mathcal{R}_{\text{OT-semantic}}. \quad (5)$$

GRPO. At each step, given a prompt $\mathbf{y}$, we sample a group of videos using SDEs $\{\mathbf{x}_0^1, \mathbf{x}_0^2, \ldots, \mathbf{x}_0^G\}$ from the video denoiser $\pi_{\theta_{\text{old}}}$, and optimize the policy model π_θ with the objective (DeepSeek-AI et al., 2026; Liu et al., 2025a):

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{\substack{\{\mathbf{x}_0^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{y}) \\ \mathbf{a}_{t,i} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{s}_{t,i})}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{T}\sum_{t=1}^{T} \max\left(-\rho_{t,i}A_i,\ -\text{clip}(\rho_{t,i}, 1-\epsilon, 1+\epsilon)\, A_i\right)\right]. \quad (6)$$

where $\rho_{t,i} = \frac{\pi_\theta(\mathbf{a}_{t,i}|\mathbf{s}_{t,i})}{\pi_{\theta_{old}}(\mathbf{a}_{t,i}|\mathbf{s}_{t,i})}$ and $A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$ is the advantage function computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ with Dual OT-aligned Rewards module.

During post-training, we use LoRA (Hu et al., 2022), freezing all parameters except the denoiser. Algorithm 1 summarizes this procedure: we generate videos using Euler ODE for direct backprop and SDE for GRPO, compute the Dual OT-aligned Rewards, and optimize the denoiser.

Algorithm 1 Post-Training with OT-aligned Rewards

Require: Pre-trained denoiser $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$; data $\mathbf{p}_{\text{data}}$; ODE solver Φ; skipping interval k; distance $d(\cdot,\cdot)$; $\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$
Ensure: $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ converges and minimizes $\mathcal{L}_{\text{total}}$
while not converged **do**
  Sample video-text $(\mathbf{x}_{\text{video}}, \mathbf{y}_{\text{text}}) \sim \mathbf{p}_{\text{data}}$, $n \sim \mathcal{U}[1, N-k]$
  $\mathbf{z}_0 \leftarrow \mathcal{E}(\mathbf{x}_{\text{video}})$ ▷ Encode $\mathbf{x}_{\text{video}}$ to latent space $\mathbf{z}_0$
  Extract text embedding $\mathbf{y}$ from $\mathbf{y}_{\text{text}}$
  $\mathbf{z}_{t_{n+k}} \sim \mathcal{N}\left(\alpha(t_{n+k})\mathbf{z}_0; \beta^2(t_{n+k})\mathbf{I}\right)$ ▷ Add noise to latent
  Perform ODE solver from $t_{n+k} \to t_n$:
  $\hat{\mathbf{z}}^{\phi}_{t_n} \leftarrow \mathbf{z}_{t_{n+k}} + (t_n - t_{n+k})\Phi(\mathbf{z}_{t_{n+k}}, t_{n+k}; \phi)$
  Compute Consistency Distillation loss:
  $\mathcal{L}_{\text{CD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \phi) \leftarrow d\left(\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}_{t_{n+k}}, t_{n+k}), \boldsymbol{g}_{\boldsymbol{\theta}^-}(\hat{\mathbf{z}}^{\phi}_{t_n}, t_n)\right)$
  Single-step ODE solver from $t_{n+k} \to 0$:
  $\hat{\mathbf{z}}^{\phi}_0 \leftarrow \mathbf{z}_{t_{n+k}} - \int_0^{t_{n+k}} \left(\gamma(t)\mathbf{z}_t + \frac{1}{2}\sigma^2(t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, \mathbf{y}, t)\right) \mathrm{d}t$
  $\hat{\mathbf{x}}_0 \leftarrow \mathcal{D}(\hat{\mathbf{z}}^{\phi}_0)$ ▷ Decode $\hat{\mathbf{z}}^{\phi}_0$ to pixel space $\hat{\mathbf{x}}_0$
  Extract video embedding $\hat{\mathbf{x}}$ from video $\hat{\mathbf{x}}_0$ and compute rewards using $\hat{\mathbf{x}}$ and $\mathbf{y}$ with OT
  $\mathcal{L}_{\text{direct}} = \mathcal{L}_{\text{CD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \phi) - \mathcal{R}_{\text{OT-quality}} - \mathcal{R}_{\text{OT-semantic}}$ or $\mathcal{L}_{\text{totalGRPO}} = \mathcal{L}_{\text{CD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \phi) + \mathcal{L}_{\text{GRPO}}$
  Backward $\mathcal{L}_{\text{direct}}$ or $\mathcal{L}_{\text{totalGRPO}}$ to update $\boldsymbol{\theta}$
  $\boldsymbol{\theta}^- \leftarrow \texttt{stop\_grad}(\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}^-)$
end while

4. Experiments

4.1. Experimental Setting

Implementation Details. We validate PISCES by post-training VideoCrafter2 (Chen et al., 2024) (2s @ 8FPS) and HunyuanVideo (Kong et al., 2025) (5s @ 25FPS), representing short- and long-video settings. We use base text-video embeddings from InternVideo2 (Wang et al., 2025a) for OT-based reward alignment. We perform post-training on $8\times$ A100 80GB GPUs for 2 days with a learning rate of $1e{-}6$, batch size 1, and accumulation 32 (direct backprop) or 4 days for GRPO (includes time for intermediate inference and visualization, actual training time is $\approx$ 30 hours for direct backpropagation and $\approx$ 78 hours for GRPO). Both the OT map $\mathbf{T}_\psi$ and critic f_ω are 3-layer MLPs with ReLU activations and LayerNorm. To train Neural OT (NOT) map, we use video-text pairs with 8-frame clips, extracted using frozen InternVideo2 (Wang et al., 2025a), and train on 1 A100 GPU for one day, equivalent to 24 A100 GPU-hours.

Datasets. Following T2V-Turbo-v2 (Li et al., 2025a), we construct a balanced dataset mixing WebVid10M (Bain et al., 2021) and VidGen-1M (Tan et al., 2024). We sample 2s clips at 8FPS for VideoCrafter2 (short-video) and 5s clips at 25FPS for HunyuanVideo (long-video). We resize frames to 512×320 for VideoCrafter2 and 848×480 for HunyuanVideo before training.

Evaluation Metrics. We evaluate PISCES with VBench (Huang et al., 2024), benchmarking T2V generation across 16 dimensions summarized into a *Quality Score* (visual fidelity, temporal coherence, *e.g.*, subject/background consistency, motion smoothness) and a *Semantic Score* (fine-grained alignment to prompts, *e.g.*, object presence, spatial relations, action correctness). The *Total Score*, a weighted sum of the two, provides a holistic measure of video fidelity and semantic alignment. We also conduct user study on 400 prompts as per VideoReward (Liu et al., 2025b).

4.2. Comparison with Existing Methods

Automatic Evaluation on VBench. Table 1 compares PISCES with existing T2V post-training methods (Li et al., 2024; 2025a; Liu et al., 2025b;d; Wang et al., 2025b) on both short-video (VideoCrafter2) and long-video (HunyuanVideo) generation. We observe that PISCES significantly outperforms both annotation-based and annotation-

Table 1. **VBench comparison on VideoCrafter2 and HunyuanVideo**. PISCES significantly outperforms existing reward-based T2V post-training methods across all scores. *Reproduced without motion guidance for fair comparison. Additional analysis in Appendix G.

Models	VideoCrafter2 (Chen et al., 2024)			HunyuanVideo (Kong et al., 2025)		
	Total	Quality	Semantic	Total	Quality	Semantic
Vanilla	80.44	82.20	73.42	83.24	85.09	75.82
VCM (Wang et al., 2023b)	73.97 ↓6.47	78.54 ↓3.66	55.66 ↓17.8	81.77 ↓1.47	84.60 ↓0.49	70.49 ↓5.33
T2V-Turbo (Li et al., 2024)	81.01 ↑0.57	82.57 ↑0.37	74.76 ↑1.34	83.86 ↑0.62	85.57 ↑0.48	77.00 ↑1.18
T2V-Turbo-v2* (Li et al., 2025a)	81.87 ↑1.43	83.26 ↑1.06	76.30 ↑2.88	84.25 ↑1.01	85.93 ↑0.84	77.52 ↑1.70
VideoReward-DPO (Liu et al., 2025b)	80.75 ↑0.31	82.11 ↓0.09	75.29 ↑1.87	83.54 ↑0.30	85.02 ↓0.07	77.63 ↑1.81
VideoDPO (Liu et al., 2025d)	81.93 ↑1.49	83.07 ↑0.87	77.38 ↑3.96	84.13 ↑0.89	85.71 ↑0.62	77.83 ↑2.01
UnifiedReward (Wang et al., 2025b)	81.43 ↑0.99	83.26 ↑1.06	74.12 ↑0.70	83.80 ↑0.56	85.46 ↑0.37	77.15 ↑1.33
PISCES	**82.75** ↑2.31	**84.05** ↑1.85	**77.54** ↑4.12	**85.45** ↑2.21	**86.73** ↑1.64	**80.33** ↑4.51

Table 2. **Automatic Evaluation on VBench**. We compare different T2V models across Quality, Semantic, and Total Scores. HunyuanVideo, post-trained with PISCES, performs the best on all scores across all models.

Metric	ModelScope (Wang et al., 2023a)	Show-1 (Zhang et al., 2024)	Pika-1.0 (Pika Labs, 2023)	Gen-3 (RunwayML, 2024)	Kling (KlingAI, 2025)	VideoCrafter2 (Chen et al., 2024)	HunyuanVideo (Kong et al., 2025)	PISCES	
								VideoCrafter2	HunyuanVideo
Quality Score	78.05	80.42	82.92	84.11	83.39	82.20	85.09	83.73	**86.73**
Semantic Score	66.54	72.98	71.77	75.17	75.68	73.42	75.82	77.63	**80.33**
Total Score	75.75	78.93	80.69	82.32	81.85	80.44	83.24	82.51	**85.45**

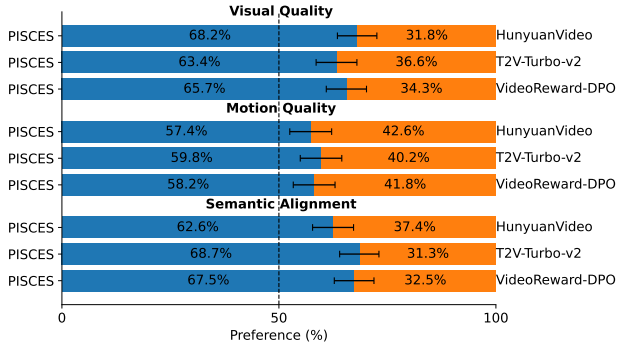

Figure 3. **Human preference study.** PISCES outperforms HunyuanVideo, T2V-Turbo-v2, VideoReward-DPO in visual quality, motion and semantic alignment, validating its effectiveness in T2V.

free approaches, achieving the highest scores across all metrics – Total, Quality, and Semantic. Table 2 provides a broader comparison with recent T2V models, both open-source (Wang et al., 2023a; Kong et al., 2025; Chen et al., 2024; Zhang et al., 2024) and closed-source (Pika Labs, 2023; KlingAI, 2025; RunwayML, 2024). We can again observe that, when post-trained with PISCES, HunyuanVideo outperforms existing T2V models on automatic VBench evaluation. This highlights the benefit of our OT formulation to align the text-video embeddings in an off-the-shelf pre-trained VLM, which strengthens supervision for the T2V post-training Rewards module. This also highlights the effectiveness of our proposed OT-aligned Quality and Semantic Rewards in performing optimal T2V post-training. We also show that PISCES's OT-aligned Rewards module is applicable to different optimizations in Appendix D.

Human Evaluation. Following VideoReward (Liu et al., 2025b), we conduct a human study on 400 prompts, evaluating three dimensions: visual quality, motion quality, and text alignment. For each prompt, we generate videos using pre-trained HunyuanVideo, post-trained T2V-Turbo-v2 (Li et al., 2025a), VideoReward-DPO (Liu et al., 2025b), and PISCES. Participants ($n = 85$) answer three questions per prompt: (1) Which video is better aligned with the text prompt? (2) Which video has better visual quality? and (3) Which video has better motion quality? As shown in Figure 3, PISCES is consistently preferred over all baselines in terms of visual quality, motion quality, and semantic alignment. This confirms that our dual OT-aligned rewards enhance both visual and semantic consistency. Furthermore, experiments on evaluation prompts from VBench (Huang et al., 2024) and VideoReward (Liu et al., 2025b) imply the generalization of our approach to out-of-domain prompts, as we train T2V models only on WebVid10M (Bain et al., 2021) and VidGen-1M (Tan et al., 2024).

Qualitative Comparison. To further assess the effectiveness of PISCES, we visually compare videos generated by PISCES against other methods in Fig. 4. Given the text prompt, we observe PISCES generates videos with improved semantic fidelity and visual coherence. Compared to baselines, PISCES is better at preserving fine-grained semantic details, such as the reflection effects on the wet pavement and the vibrant color contrast in the scene (w.r.t. Semantic Score in VBench). The generated subject is also more globally consistent across frames, reducing temporal flicker and maintaining a stable appearance (w.r.t. to Quality Score in VBench). These results align with our quantitative findings from the previous section, validating the effectiveness of Dual OT-aligned Rewards to enhance both visual quality and text-video alignment in T2V models.

4.3. Ablation Study

Effectiveness of OT in Text-Video Alignment. Table 3 (Rows 2 and 5) compares PISCES with and without OT on VideoCrafter2, isolating the effect of aligning text and video embedding distributions via OT. For the Quality Reward, we replace the OT-aligned $\mathbf{T}^\star(\mathbf{y}_{[\text{CLS}]})$ in Eq. (2) with the raw $\mathbf{y}_{[\text{CLS}]}$, and for the Semantic Reward, we remove the POT–guided attention $\tilde{\mathbf{A}}$ in Eq. (4) and use vanilla cross-

Figure 4. Qualitative comparison of T2V models. PISCES produces videos with better semantic fidelity and visual quality, accurately capturing key details such as the reflective wet pavement and vibrant neon lighting.

Table 3. **Ablation Study.** OT alignment improves both Quality and Semantic Scores, while Quality Reward enhances visual quality and Semantic Reward improves text-video correspondence. Full PISCES achieves the **best** performance.

Method	OT	$\mathcal{R}_{\text{semantic}}$	$\mathcal{R}_{\text{quality}}$	Total Score	Quality Score	Semantic Score
Vanilla (VideoCrafter2)	✗	✗	✗	80.44	82.20	73.42
PISCES w/o OT	✗	✓	✓	81.92	83.44	75.82
PISCES + $\mathcal{R}_{\text{OT-quality}}$	✓	✗	✓	<u>82.21</u>	**83.77**	75.97
PISCES + $\mathcal{R}_{\text{OT-semantic}}$	✓	✓	✗	81.70	82.87	<u>76.99</u>
PISCES Full	✓	✓	✓	**82.51**	<u>83.73</u>	**77.63**

attention **A**. We find that OT is critical for improving both Quality and Semantic performance. PISCES attains a Semantic Score of 77.63, outperforming the variant without OT (75.82), demonstrating that distribution and token-level alignment prior to reward computation substantially enhances semantic correspondence. OT also improves Quality Score (83.73 vs. 83.44), confirming its role in structuring the feature space for more effective T2V post-training. Overall, these results show that OT-aligned embeddings provide stronger supervision than off-the-shelf pre-trained VLM embeddings used in existing post-training methods.

Impact of OT-aligned Quality and Semantic Rewards. Table 3 analyzes the contribution of each OT-aligned reward. Using only the Quality Reward (Row 3) improves the VBench Quality Score from 82.20 to 83.77 (Row 1), demonstrating its effectiveness in enhancing global coherence and visual quality. In contrast, using only the Semantic Reward (Row 4) substantially increases the Semantic Score from 73.42 to 76.99, highlighting its ability to capture fine-grained text–video alignment such as object presence and actions. Combining both rewards (Row 5) yields the best overall performance, improving both Quality and Semantic Scores over single-reward variants. This confirms their complementary roles within PISCES 's OT-aligned space. For results across all 16 VBench dimensions, see Appendix E.

Figure 5. Cross-attention maps (left) are diffuse, OT plan without spatio-temporal constraints (middle) misaligns tokens, while our constrained OT plan (right) produces accurate correspondences.

Table 4. Comparison of alignment methods. OT improves text-video alignment (higher Mutual KNN) and preserves text embedding structure (higher Spearman Correlation). Post-training with OT-aligned rewards achieves the best Quality and Semantic Scores.

Method	Mutual KNN ↑	Spearman Correlation ↑	Quality Score ↑	Semantic Score ↑
Contrastive	0.2135	-	85.57	77.00
Mapping w/ L2	0.2318	0.4873	84.89	77.12
Mapping w/ KL	0.2284	0.4720	84.72	77.15
OT (PISCES)	**0.2597**	**0.9018**	**86.73**	**80.33**

4.4. Optimal Transport Analysis

To validate the effectiveness of OT (Villani, 2009; Cuturi, 2013) in aligning text and video embeddings, we conduct both qualitative and quantitative analyses. We extract $10,000$ text-video pairs from WebVid10M (Bain et al., 2021) using pre-trained VLM InternVideo2.

OT Plan. Figure 5 compares standard cross-attention maps with our token-level OT plans. While cross-attention alone produces diffuse activations, and unconstrained OT plans misalign tokens, incorporating spatio-temporal constraints in the OT cost matrix yields meaningful correspondences. This highlights the benefit of discrete OT in our Semantic Reward: it ensures fine-grained alignment of text tokens with semantically and spatio-temporally consistent video regions, directly improving localized supervision during post-training. We further validate that our designed discrete POT helps improve the video-text matching performance of pre-trained InternVideo2 by 8.11% in Appendix F.

Quantitative Analysis. Mutual KNN (Huh et al., 2024) measures cross-modal alignment by evaluating k-nearest-neighbor overlap between text and video embeddings, where higher values indicate stronger alignment. Spearman correlation r assesses structural preservation by measuring rank consistency before and after alignment. As shown in Table 4, OT achieves the highest Mutual KNN and Spearman correlation among all methods, indicating effective distribution alignment with minimal structural distortion. Post-training HunyuanVideo with OT-aligned rewards further yields the best Quality and Semantic Scores, confirming improvements in both visual fidelity and text–video consistency.

OT Map. Fig. 6 (left) presents a t-SNE projection of text and video embeddings. The original text embeddings (blue) are largely misaligned with video embeddings (orange), highlighting distributional gaps in VLMs (leading to sub-optimal text-video alignment). OT-transformed text embeddings (green) shift significantly closer to video embeddings, demonstrating improved alignment. Fig. 6 (right) further supports this observation via pairwise distance distribution. The distribution of text embeddings after OT transformation closely resembles the original one, confirming OT aligns embeddings without distorting their internal relationships.

Impact on T2V Post-Training. In Figure 7, we conduct a controlled experiment using the prompt *"A robot reaching for a red/green apple on the table"* with the same random seed. Post-training with OT-aligned rewards (left) maintains structural consistency – background, robot appearance, and stable motion, with only the apple color changing as expected. In contrast, L2 loss mapping (right) produces unstable outputs: robot appearance and object placement vary unpredictably, and artifacts such as disappearing objects emerge. These results confirm that distorted text embeddings distribution harms reward-based post-training, leading to unstable sampling and degraded video quality. Overall, these findings underscore the advantages of OT alignment. By ensuring rewards operate on a well-structured space, OT prevents distributional distortions, resulting in improved sampling stability and higher-quality T2V generation.

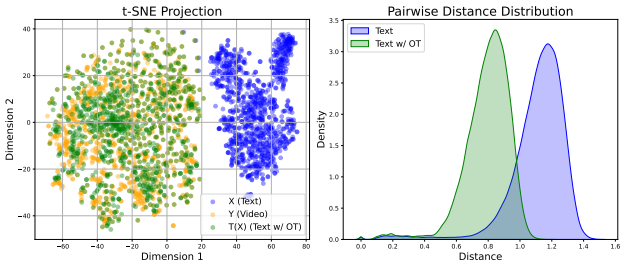

Figure 6. t-SNE shows OT aligns text embeddings (green) closer to video embeddings distribution (orange). Pairwise distance distribution indicates OT preserves text embedding structure.

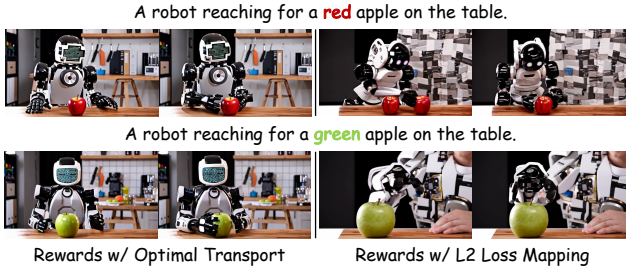

Figure 7. Post-training with OT-aligned rewards (left) yields consistent outputs with expected color changes, while L2 loss mapping (right) causes sampling inconsistencies and visual artifacts.

5. Conclusion

We introduce PISCES, the first annotation-free post-training T2V algorithm that outperforms all existing annotation-based and annotation-free methods on VBench and human evaluation. Overcoming the limitation of existing annotation-free methods, which rely on VLM embeddings misaligned at both distributional and token levels, PISCES introduces a novel Dual OT-aligned Rewards module through the lens of OT. It comprises a Distributional OT-aligned Quality Reward and a Discrete token-level Semantic Reward to significantly improve visual quality and semantic consistency across short- and long-video generation. PISCES paves the way for scalable, principled post-training in T2V and offers a general blueprint for OT-based reward design in multimodal generation.

Acknowledgments

This work was carried out during Minh-Quan's internship at Microsoft. Dimitris Samaras was supported in part by NSF grants IIS-2123920 and IIS-2212046. Xianfeng David Gu was supported by NIH R21EB029733.

Impact Statement

This work presents PISCES, a scalable and annotation-free post-training framework for improving text-to-video (T2V) generation via Dual OT-aligned Rewards. By reducing reliance on costly human annotations, PISCES offers a principled alternative to learning reward models from human preferences. Its distributional OT alignment strengthens global text–video coherence and visual quality, while its token-level OT alignment improves fine-grained semantic grounding, together enhancing both visual fidelity and semantic consistency. These improvements may broaden the usability of T2V models for beneficial applications such as education (visual explanations and instructional content), robotics and embodied AI (rapid prototyping of scenarios and simulations), and scientific visualization (communicating complex processes via controllable video synthesis).

More broadly, by making reward-based post-training more scalable, our approach enables researchers and practitioners to iterate on alignment methods without requiring large-scale preference datasets, potentially lowering barriers to experimentation and enabling wider participation. Because PISCES is annotation-free, it reduces the dependence on extensive human labeling pipelines, helping mitigate practical challenges related to cost, scalability, and the need to expose annotators to sensitive content.

PISCES is a training framework that improves alignment and quality using existing model signals, and it does not introduce any new categories of risk beyond those already associated with modern T2V systems. We emphasize that responsible deployment is important: applications should incorporate established safeguards (e.g., content moderation policies, provenance, or disclosure mechanisms when appropriate, and careful dataset and evaluation practices). We encourage future work to continue improving mechanisms for the safe deployment of T2V systems, particularly in high-impact or public-facing settings.

References

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., and Shan, Y. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Learning universal image-text representations. In *European Conference on Computer Vision (ECCV)*, pp. 104–120, 2020.

Chung, H. W., Garcia, X., Roberts, A., Tay, Y., Firat, O., Narang, S., and Constant, N. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kXwdL1cWOAi.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z.,

Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2026. URL https://arxiv.org/abs/2501.12948.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

GLM, T., :, Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., Zhang, J., Sun, J., Li, J., Zhao, L., Wu, L., Zhong, L., Liu, M., Huang, M., Zhang, P., Zheng, Q., Lu, R., Duan, S., Zhang, S., Cao, S., Yang, S., Tam, W. L., Zhao, W., Liu, X., Xia, X., Zhang, X., Gu, X., Lv, X., Liu, X., Liu, X., Yang, X., Song, X., Zhang, X., An, Y., Xu, Y., Niu, Y., Yang, Y., Li, Y., Bai, Y., Dong, Y., Qi, Z., Wang, Z., Yang, Z., Du, Z., Hou, Z., and Wang, Z. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Han, H., Zheng, Q., Dai, G., Luo, M., and Wang, J. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26679–26688, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Izquierdo, S. and Civera, J. Optimal transport aggregation for visual place recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 17658–17668, 2024.

Katageri, S., De, A., Devaguptapu, C., Prasad, V., Sharma, C., and Kaul, M. Synergizing contrastive learning and optimal transport for 3d point cloud domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2942–2951, 2024.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

KlingAI. Klingai - ai-powered solutions, 2025. URL https://www.klingai.com.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., and Zhong, C. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Le, M.-Q., Mittal, G., Meng, T., Iftekhar, A. S. M., Suryanarayanan, V., Patra, B., Samaras, D., and Chen, M. Hummingbird: High fidelity image generation via multimodal context alignment. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=6kPBThI6ZJ.

Le, M.-Q., Zhu, Y., Kalogeiton, V., and Samaras, D. What about gravity in video generation? post-training newton's laws with verifiable rewards, 2025b. URL https://arxiv.org/abs/2512.00425.

Li, J., Feng, W., Fu, T.-J., Wang, X., Basu, S., Chen, W., and Wang, W. Y. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In *Advances in Neural Information Processing Systems*, 2024.

Li, J., Long, Q., Zheng, J., Gao, X., Piramuthu, R., Chen, W., and Wang, W. Y. T2v-turbo-v2: Enhancing video

model post-training through data, reward, and conditional guidance design. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=BZwXMqu4zG.

Li, Z., Li, S., Wang, Z., Lei, N., Luo, Z., and Gu, D. X. Dpm-ot: a new diffusion probabilistic model based on optimal transport. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 22624–22633, 2023.

Li, Z., Wang, W., Zhao, Y., Li, W., Lei, N., and Gu, X. Hyper-spherical optimal transport for semantic alignment in text-to-3d end-to-end generation. *IEEE Transactions on Visualization and Computer Graphics*, 31(10), 2025b.

Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., ZHANG, D., and Ouyang, W. Flow-GRPO: Training flow matching models via online RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=oCBKGw5HNf.

Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., Liu, X., Yang, F., Wan, P., ZHANG, D., Gai, K., Yang, Y., and Ouyang, W. Improving video generation with human feedback. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=nHkg4yc7SP.

Liu, M., Wang, L., Zhou, S., Xia, K., Sun, X., and Hua, G. Boosting point-supervised temporal action localization through integrating query reformation and optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13865–13875, June 2025c.

Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., and Chen, Q. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 8009–8019, June 2025d.

Lu, C. and Song, Y. Simplifying, stabilizing and scaling continuous-time consistency models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LyJi5ugyJx.

Ma, Y., Wu, X., Sun, K., and Li, H. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, October 2025.

Pika Labs. Pika 1.0: Video generation model. https://pika.art, 2023. Platform announcement; proprietary model.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

RunwayML. Introducing gen-3 alpha: A new frontier for video generation, 2024. URL https://runwayml.com/research/introducing-gen-3-alpha.

Seaweed, T., Yang, C., Lin, Z., Zhao, Y., Lin, S., Ma, Z., Guo, H., Chen, H., Qi, L., Wang, S., Cheng, F., Zuo, F., Zeng, X., Yang, Z., Kong, F., Wei, M., Qing, Z., Xiao, F., Hoang, T., Zhang, S., Zhu, P., Zhao, Q., Yan, J., Gui, L., Bi, S., Li, J., Ren, Y., Wang, R., Li, H., Xiao, X., Liu, S., Ling, F., Zhang, H., Wei, H., Kuang, H., Duncan, J., Zhang, J., Zheng, J., Sun, L., Zhang, M., Sun, R., Zhuang, X., Li, X., Xia, X., Chi, X., Peng, Y., Wang, Y., Wang, Y., Zhao, Z., Chen, Z., Song, Z., Yang, Z., Feng, J., Yang, J., and Jiang, L. Seaweed-7b: Cost-effective training of video generation foundation model, 2025. URL https://arxiv.org/abs/2504.08685.

Shi, L., Fan, J., and Yan, J. OT-CLIP: Understanding and generalizing CLIP via optimal transport. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=X8uQ1TslUc.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *ICML*, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Sun, X., Chen, Y., Huang, Y., Xie, R., Zhu, J., Zhang, K., Li, S., Yang, Z., Han, J., Shu, X., Bu, J., Chen, Z., Huang, X., Lian, F., Yang, S., Yan, J., Zeng, Y., Ren, X., Yu, C., Wu, L., Mao, Y., Xia, J., Yang, T., Zheng, S., Wu, K., Jiao, D., Xue, J., Zhang, X., Wu, D., Liu, K., Wu, D., Xu, G., Chen, S., Chen, S., Feng, X., Hong, Y., Zheng, J., Xu, C., Li, Z., Kuang, X., Hu, J., Chen, Y., Deng, Y., Li, G., Liu, A., Zhang, C., Hu, S., Zhao, Z., Wu, Z., Ding, Y., Wang, W., Liu, H., Wang, R., Fei, H., Yu, P., Zhao, Z., Cao, X.,

Wang, H., Xiang, F., Huang, M., Xiong, Z., Hu, B., Hou, X., Jiang, L., Ma, J., Wu, J., Deng, Y., Shen, Y., Wang, Q., Liu, W., Liu, J., Chen, M., Dong, L., Jia, W., Chen, H., Liu, F., Yuan, R., Xu, H., Yan, Z., Cao, T., Hu, Z., Feng, X., Du, D., Yu, T., Tao, Y., Zhang, F., Zhu, J., Xu, C., Li, X., Zha, C., Ouyang, W., Xia, Y., Li, X., He, Z., Chen, R., Song, J., Chen, R., Jiang, F., Zhao, C., Wang, B., Gong, H., Gan, R., Hu, W., Kang, Z., Yang, Y., Liu, Y., Wang, D., and Jiang, J. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent, 2024. URL https://arxiv.org/abs/2411.02265.

Tan, Z., Yang, X., Qin, L., and Li, H. Vidgen-1m: A large-scale dataset for text-to-video generation, 2024. URL https://arxiv.org/abs/2408.02629.

Tong, A., FATRAS, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Villani, C. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., and Naik, N. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, June 2024.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., and Zhang, S. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., and Sang, N. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023b.

Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., He, C., Luo, P., Liu, Z., Wang, Y., Wang, L., and Qiao, Y. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2307.06942.

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., Jiang, T., Li, S., Xu, J., Zhang, H., Huang, Y., Qiao, Y., Wang, Y., and Wang, L. Internvideo2: Scaling foundation models for multimodal video understanding. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (eds.), *Computer Vision – ECCV 2024*, pp. 396–416, Cham, 2025a. Springer Nature Switzerland.

Wang, Y., Zang, Y., Li, H., Jin, C., and Wang, J. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.

Xie, Y., Zeng, Z., Zhang, H., Ding, Y., Wang, Y., Wang, Z., Chen, B., and Liu, H. Discovering fine-grained visual-concept relations by disentangled optimal transport concept bottleneck models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 30199–30209, June 2025.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Yang, X., Yang, M., JIA, G., Qin, L., Tan, Z., and Li, H. Dual-IPO: Dual-iterative preference optimization for text-to-video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mu8sO1Vw0C.

Yuan, H., Zhang, S., Wang, X., Wei, Y., Feng, T., Pan, Y., Zhang, Y., Liu, Z., Albanie, S., and Ni, D. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6463–6474, June 2024.

Zhang, D. J., Wu, J. Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., Gao, D., and Shou, M. Z. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pp. 1–15, 2024.

Appendix

In this technical appendix, we provide additional details, ablations, and analyses that support the main findings of our work. Section A introduces the Consistency Distillation (CD) (Song et al., 2023; Wang et al., 2023b; Lu & Song, 2025) mechanism used to efficiently integrate OT-aligned rewards into the PISCES post-training process. Section B describes the Neural Optimal Transport (NOT) (Villani, 2009; Korotin et al., 2023) formulation for aligning the distributions of text and video embeddings, while Section C presents our discrete token-level OT optimization via the entropic unbalanced Sinkhorn algorithm (Cuturi, 2013).

Section D shows that our dual reward module is compatible with different optimization paradigms (direct backpropagation and GRPO (DeepSeek-AI et al., 2026; Liu et al., 2025a)). Section E provides a detailed ablation study on the impact of OT alignment and reward types across the full 16 dimensions of VBench (Huang et al., 2024). Section F quantifies the impact of partial OT and structured spatio-temporal constraints on matching accuracy. Section G studies the role of motion guidance during post-training.

Section H evaluates PISCES under ViCLIP-based (Wang et al., 2024) reward models to confirm generalizability across video-text encoders. Section I reports alignment performance on out-of-distribution prompts. Section J presents a hyperparameter sweep over the OT loss weights (γ, η) to assess sensitivity. Section K highlights a failure case of partial OT due to base encoder limitations. Section L reports inter-rater reliability and checks for category-level bias in human evaluation. Section M discusses an adaptive weighting strategy for reward fusion. Section N assesses the interaction of the dual OT-aligned rewards during training. Section O analyzes reward hacking risks and shows how CD loss mitigates them. Finally, Section P reports GPU-hour cost and runtime efficiency compared to baseline methods.

A. Consistency Distillation

Consistency Models (CMs) (Song et al., 2023; Wang et al., 2023b; Lu & Song, 2025) improve efficiency by enforcing self-consistency in the PF-ODE trajectory. A learned function $\boldsymbol{g} : (\mathbf{z}_t, t) \mapsto \mathbf{z}_\epsilon$ satisfies $\boldsymbol{g}(\mathbf{z}_t, t) = \boldsymbol{g}(\mathbf{z}_{t'}, t'), \quad \forall t, t' \in [\epsilon, T]$. A pre-trained diffusion model is distilled into a CM via the Consistency Distillation loss:

$$\mathcal{L}_{\mathrm{CD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \phi) = \mathbb{E}_{\mathbf{z},t}\left[d\left(\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}_{t+k}, t+k), \boldsymbol{g}_{\theta^-}(\hat{\mathbf{z}}^{\phi}_{t_n}, t_n)\right)\right], \tag{7}$$

where an ODE solver Φ estimates $\hat{\mathbf{z}}^{\phi}_{t_n} \leftarrow \mathbf{z}_{t_{n+k}} + (t_n - t_{n+k})\Phi(\mathbf{z}_{t_{n+k}}, t_{n+k}; \phi)$. To stabilize learning, EMA updates $\boldsymbol{\theta}^- \leftarrow \texttt{stop\_grad}(\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}^-)$. Consistency distillation allows reward fine-tuning to backpropagate via single-step denoising, approximating multi-step denoising to enhance both efficiency and supervision signal.

B. Distributional Alignment with OT

We first address the embeddings distribution misalignment in pre-trained VLMs by formulating the alignment as an OT problem. Specifically, given text embeddings $\mathcal{Y}$ and real video embeddings $\mathcal{X}$ extracted from a pre-trained VLM, we learn an OT map $\mathbf{T} : \mathcal{Y} \to \mathcal{X}$ using NOT (Korotin et al., 2023). This OT mapping transforms text embeddings into a semantically-aligned space with video embeddings, significantly reducing the inherent distributional mismatch (see Figure 2 left). We define the OT problem as:

$$\sup_{f} \inf_{\mathbf{T}} \int_{\mathcal{X}} f(\mathbf{x}) d\nu(\mathbf{x}) + \int_{\mathcal{Y}} \left(\mathbf{c}(\mathbf{y}, \mathbf{T}(\mathbf{y})) - f(\mathbf{T}(\mathbf{y}))\right) d\mu(\mathbf{y}), \tag{8}$$

where the cost function is the squared Euclidean distance, $\mathbf{c}(\mathbf{y}, \mathbf{x}) = \|\mathbf{y} - \mathbf{x}\|^2$. We implement this via iterative optimization of the transport map $\mathbf{T}_\psi$ and potential function f_ω parameterized by neural networks, as shown in Algorithm 2. The resulting OT-aligned embeddings $\mathbf{T}^\star(\mathbf{y})$ preserve original embedding structure while aligning distributions.

Algorithm 2 Text-Video Embeddings Distribution Alignment w/ OT

Require: text and video embedding distributions $\boldsymbol{\mu}, \boldsymbol{\nu}$; mapping network $\mathbf{T}_\psi : \mathcal{Y} \to \mathcal{X}$;
potential network $f_\omega : \mathcal{X} \to \mathbb{R}$; number of inner iterations K_T;
cost function $\mathbf{c} : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$
Ensure: learned stochastic OT map $\mathbf{T}_\psi$ representing an OT plan between distributions $\boldsymbol{\mu}, \boldsymbol{\nu}$
while not converged **do**
  $\texttt{unfreeze}(\mathbf{T}_\psi)$; $\texttt{freeze}(f_\omega)$ ▷ $\mathbf{T}$ optimization
  for $k_T = 1, 2, \ldots, K_T$ **do**
    Sample a batch of text embeddings $Y \sim \boldsymbol{\mu}$
    $\mathcal{L}_{\mathbf{T}} \leftarrow \frac{1}{|Y|}\sum_{\mathbf{y}\in Y}\left[\mathbf{c}(\mathbf{y}, \mathbf{T}_\psi(\mathbf{y})) - f_\omega(\mathbf{T}_\psi(\mathbf{y}))\right]$
    Backward $\mathcal{L}_{\mathbf{T}}$ and update ψ using $\frac{\partial \mathcal{L}_{\mathbf{T}}}{\partial \psi}$
  end for
  $\texttt{freeze}(\mathbf{T}_\psi)$; $\texttt{unfreeze}(f_\omega)$ ▷ f optimization
  Sample batch of video and text embeddings $X \sim \boldsymbol{\nu}, Y \sim \boldsymbol{\mu}$
  $\mathcal{L}_f \leftarrow \frac{1}{|Y|}\sum_{\mathbf{y}\in\mathcal{Y}} f_\omega(\mathbf{T}_\psi(\mathbf{y})) - \frac{1}{|X|}\sum_{\mathbf{x}\in X} f_\omega(\mathbf{x})$
  Backward $\mathcal{L}_f$ and update ω using $\frac{\partial \mathcal{L}_f}{\partial \omega}$
end while

C. Discrete OT for Semantic Alignment

Problem setup. Given text tokens $\{\mathbf{y}_i\}_{i=1}^N$ and video patch tokens $\{\mathbf{x}_j\}_{j=1}^M$ from a cross-attention layer, let $\mathbf{A} \in \mathbb{R}_+^{N\times M}$ denote the vanilla attention (row-normalized), t_j the frame index of patch j, and $s_j \in \mathbb{R}^2$ its spatial coordinate on a $h \times w$ grid. We construct a spatio-temporal, semantics-aware cost matrix $\mathbf{C} \in \mathbb{R}_+^{N\times M}$ as

$$\mathbf{C}_{ij} = \underbrace{1 - \cos(\mathbf{y}_i, \mathbf{x}_j)}_{\text{semantic}} + \gamma \underbrace{\left|\tau(\mathbf{y}_i) - t_j\right|}_{\text{temporal}} + \eta \underbrace{\left\|\pi(\mathbf{y}_i) - s_j\right\|_2}_{\text{spatial}},$$

where $\tau(\mathbf{y}_i) = \sum_k A_{ik}\, t_k$ and $\pi(\mathbf{y}_i) = \sum_k A_{ik}\, s_k$ are attention-weighted expectations of frame index and spatial

position, respectively. Each component is range-normalized to $[0, 1]$, followed by a min–max normalization of $\mathbf{C}$ to $[0, 1]$ for numerical stability.

Partial entropic OT. Let uniform marginals $\boldsymbol{\mu} = \frac{1}{N}\mathbf{1}_N$ and $\boldsymbol{\nu} = \frac{1}{M}\mathbf{1}_M$. We solve a *partial* OT problem via an entropic, unbalanced Sinkhorn objective:

$$\min_{\mathbf{P}\geq 0} \ \langle \mathbf{P}, \mathbf{C}\rangle + \epsilon \sum_{i,j} \mathbf{P}_{ij}(\log \mathbf{P}_{ij} - 1)$$
$$\text{s.t. } \mathbf{P}\mathbf{1}_M \approx \tau_a \boldsymbol{\mu}, \qquad \mathbf{P}^\top \mathbf{1}_N \approx \tau_b \boldsymbol{\nu}.$$

where $\epsilon > 0$ is the entropic temperature and $(\tau_a, \tau_b) \in (0, 1]$ relax the marginals to achieve an effective transported fraction $m \in (0, 1]$ (we use $m = 0.9$). This yields a soft plan $\mathbf{P}^\star$ that *selectively* matches informative text tokens to consistent video regions, avoiding over-forced alignments. Algorithm 3 describes our detailed implementation of solving partial OT using entropic (unbalanced) Sinkhorn.

Attention fusion (structure prior). We fuse $\mathbf{P}^\star$ with vanilla attention $\mathbf{A}$ in log-space:

$$\tilde{\mathbf{A}} \ \propto \ \exp\Big(\log(\mathbf{A} + \varepsilon) \ + \ \log(\mathbf{P}^\star + \varepsilon)\Big),$$

with small $\varepsilon > 0$ for stability. Gradients flow through $\mathbf{A}$ while $\mathbf{P}^\star$ acts as a detached structural prior.

Semantic reward. Let VTM be the pre-trained video–text matching head. Using $\tilde{\mathbf{A}}$ to aggregate patch features, the Semantic Alignment Reward is

$$\mathcal{R}_{\text{OT-semantic}} \ = \ \text{softmax}\Big(\texttt{VTM}\big[\tilde{\mathbf{A}} \cdot \hat{\mathbf{x}}\big]\Big)_{(\text{idx}=1)}.$$

In practice, POT is applied per head and per cross-attention layer. We set $(\gamma, \eta) = (0.2, 0.2)$, $\epsilon = 0.05$, and $m = 0.9$. Please refer to Appendix F for hyperparameter selection and ablation study.

D. Automatic Evaluation with Different Optimization Paradigms

We optimize the same OT-aligned reward under two training routes: (i) direct backpropagation through the reward models and (ii) RL fine-tuning via GRPO. As reported in Table 5, both procedures yield consistent gains over the vanilla baselines on VBench (Total/Quality/Semantic) for VideoCrafter2 and HunyuanVideo, and the resulting scores are comparable across paradigms. This indicates that our reward provides meaningful supervision signals whose benefits are largely agnostic to the optimization routine.

E. Comprehensive Ablation Study

Effectiveness of OT Alignment. Comparing PISCES w/o OT (which post-trains with pre-trained VLM embeddings) against full PISCES in Table 6, we observe significant improvements in both Quality Score ($83.44 \rightarrow 83.73$) and Semantic Score ($75.82 \rightarrow 77.63$). This confirms that aligning text-video distributions before post-training enhances both global coherence and fine-grained text-video correspondence.

Algorithm 3 Partial OT via Entropic (Unbalanced) Sinkhorn

Require: Cost matrix $\mathbf{C} \in \mathbb{R}_+^{N\times M}$ (normalized to $[0, 1]$), entropic temperature $\epsilon > 0$, target transported fraction $m \in (0, 1]$, max iterations K, tolerance δ
Ensure: Transport plan $\mathbf{P}^\star \in \mathbb{R}_+^{N\times M}$
Uniform marginals: $\boldsymbol{\mu} = \frac{1}{N}\mathbf{1}_N, \ \boldsymbol{\nu} = \frac{1}{M}\mathbf{1}_M$
Map partial mass to unbalanced strength:

$$\rho \leftarrow \begin{cases} \infty & \text{if } m \geq 0.999 \\ \epsilon \cdot \frac{m}{1-m} & \text{otherwise} \end{cases},$$

$$\tau(\rho) \leftarrow \begin{cases} 1 & \text{if } \rho = \infty \\ \frac{\rho}{\rho+\epsilon} & \text{else} \end{cases}$$

Set relaxations: $\tau_a \leftarrow \tau(\rho), \ \tau_b \leftarrow \tau(\rho)$
Log-kernel: $\log \mathbf{K} \leftarrow -\mathbf{C}/\epsilon$
Initialize $\log \mathbf{u} \leftarrow \mathbf{0}_N, \ \log \mathbf{v} \leftarrow \mathbf{0}_M; \ \log \boldsymbol{\mu} \leftarrow \log(\boldsymbol{\mu}), \log \boldsymbol{\nu} \leftarrow \log(\boldsymbol{\nu})$
for $k = 1$ to K **do**
  $\log(\mathbf{K}\mathbf{v}) \leftarrow \log \sum_j \exp\big(\log \mathbf{K}_{:,j} + \log \mathbf{v}_j\big)$
  $\log \mathbf{u}_{\text{new}} \leftarrow \tau_a \cdot \big(\log \boldsymbol{\mu} - \log(\mathbf{K}\mathbf{v})\big)$
  $\log(\mathbf{K}^\top \mathbf{u}_{\text{new}}) \leftarrow \log \sum_i \exp\big(\log \mathbf{K}_{i,:} + \log \mathbf{u}_{\text{new},i}\big)$
  $\log \mathbf{v}_{\text{new}} \leftarrow \tau_b \cdot \big(\log \boldsymbol{\nu} - \log(\mathbf{K}^\top \mathbf{u}_{\text{new}})\big)$
  if $\max\{\|\log \mathbf{u}_{\text{new}} - \log \mathbf{u}\|_\infty, \ \|\log \mathbf{v}_{\text{new}} - \log \mathbf{v}\|_\infty\} < \delta$ **then**
    break
  end if
  $\log \mathbf{u} \leftarrow \log \mathbf{u}_{\text{new}}, \quad \log \mathbf{v} \leftarrow \log \mathbf{v}_{\text{new}}$
end for
Recover plan: $\log \mathbf{P} \leftarrow \log \mathbf{u}\mathbf{1}_M^\top \ + \ \log \mathbf{K} \ + \ \mathbf{1}_N \log \mathbf{v}^\top$
$\mathbf{P}^\star \leftarrow \exp(\log \mathbf{P})$

Table 5. **Automatic VBench comparison on VideoCrafter2 and HunyuanVideo**. PISCES post-training with direct backpropagation or RL fine-tuning GRPO achieves comparable performance and shows strong improvement over the pre-trained models.

Models	VideoCrafter2 (Chen et al., 2024)			HunyuanVideo (Kong et al., 2025)		
	Total	Quality	Semantic	Total	Quality	Semantic
Vanilla	80.44	82.20	73.42	83.24	85.09	75.82
PISCES (Direct Backpropagation)	82.51 ↑2.07	83.73 ↑1.53	**77.63** ↑4.21	85.05 ↑1.81	**86.84** ↑1.75	77.89 ↑2.07
PISCES (GRPO)	**82.75** ↑2.31	**84.05** ↑1.85	77.54 ↑4.12	**85.45** ↑2.21	86.73 ↑1.64	**80.33** ↑4.51

Impact of Quality and Semantic Alignment Rewards. To assess the individual effects of OT-aligned Quality and Semantic Reward, we compare PISCES w/ $\mathcal{R}_{\text{OT-quality}}$ and $\mathcal{R}_{\text{OT-semantic}}$ separately. Quality Reward primarily improves global coherence, reflected in gains in Quality Score (83.77), Aesthetic Quality (66.92), and Subject Consistency (97.07). Meanwhile, Semantic Reward enhances fine-grained alignment, leading to improvements in Semantic Score (76.99), Human Action (96.60), and Spatial Relation (44.97).

Full PISCES. Integrating both rewards (full PISCES) results in the highest Total Score (82.51). Notably, Overall Consistency (29.10) and Temporal Style (26.97) also im-

prove, reinforcing that the combination of OT alignment and both rewards provides the best optimization signal for text-video post-training.

F. Optimal Transport Plan Analysis

Table 7 analyzes the effect of Partial OT and spatio-temporal constraints on video-text matching within InternVideo2. We evaluate on 10,000 video-text pairs sampled from WebVid10M (Bain et al., 2021).

We observe that using Partial OT with a mass parameter $m = 0.9$ achieves the best score of **89.36%**, improving by **+8.11%** over vanilla cross-attention. This indicates that not all text tokens need to be matched to visual patches. For example, uninformative words (*e.g.*, articles or stopwords) need not be explicitly grounded in the visual domain. Allowing partial transport filters out such noisy matches while preserving key semantic correspondences. Conversely, setting $m = 0.5$ removes too many tokens, causing essential words to be ignored and leading to degraded alignment.

Regarding constraints, our designed cost function with both spatial ($\eta = 0.2$) and temporal ($\gamma = 0.2$) penalties yields the highest performance, boosting video-text matching by **8.11%** (from 81.25% to 89.36%). This demonstrates that integrating spatio-temporal structure into OT provides sharper and more accurate token-level correspondences, thereby enhancing fine-grained text-video alignment. Overall, these results confirm the effectiveness of Partial OT with structured constraints in improving alignment quality.

G. Motion Guidance in T2V Post-Training

For a fair comparison to highlight the impact of rewards, we provide addition ablations (see Table 8) where we evaluate both methods with and without motion guidance. Without motion guidance, T2V-Turbo-v2 (Li et al., 2025a) performance drops to 83.26 (Quality) and 76.30 (Semantic) on VideoCrafter2 (Chen et al., 2024). In comparison, PISCES-direct still achieves stronger quality (83.73, +0.47) and significantly higher semantic alignment (77.63, +1.33), highlighting the effectiveness of our OT-aligned rewards. Conversely, by adding motion guidance to PISCES-direct makes it outperform T2V-Turbo-v2 (Li et al., 2025a) across all metrics on VideoCrafter2 (Chen et al., 2024).

More notably, on HunyuanVideo (Kong et al., 2025), PISCES-direct *without motion guidance* already **outperforms T2V-Turbo-v2 with motion** in all metrics (85.05 vs 84.50 Total Score), and this advantage is also reflected in our human preference study (Figure 3). This clearly demonstrates that our reward formulation, rather than the optimization strategy or motion module, is the key driver behind the improvements.

H. PISCES with ViCLIP Evaluator

To test whether PISCES improvements generalize beyond InternVideo2 (Wang et al., 2025a), we conducted a controlled experiment where both our rewards (Distributional OT and Semantic OT) are based on ViCLIP (Wang et al., 2024), a CLIP-based video–text encoder that differs from InternVideo2 in training corpus and representation space. The results in Table 9 confirm that PISCES remains effective even when rewards and evaluation use ViCLIP (Wang et al., 2024), achieving comparable performance across both benchmarks and further mitigating concerns of reward overfitting to a specific model family. These consistent results show the generality of our OT-aligned reward formulation and demonstrate that the observed improvements are not confined to InternVideo2 features.

I. Scope of Generalization

To assess the robustness of PISCES beyond the WebVid10M (Bain et al., 2021) and VidGen-1M (Tan et al., 2024) domain, we curated 100 diverse out-of-distribution (OOD) prompts. These prompts cover challenging and underrepresented scenarios including robotics actions, embodied tasks, procedural instructions, abstract concepts, and rare object-event compositions. Example prompts include:

- *A robot arm with a red gripper picks up a blue cube and sorts it into a green bin on a moving conveyor belt in a bright factory hall.*
- *A tiny mechanical mouse navigates through a labyrinth of gears inside a giant clock, camera close-up on its delicate paws.*

To test alignment under these harder OOD conditions, we compare the cosine similarity between text prompt embeddings and generated video embeddings using the ViCLIP (Wang et al., 2024) encoder. Results in Table 10 confirm that PISCES exhibits stronger alignment even under distribution shifts and when evaluated using an independent video-text encoder ViCLIP (Wang et al., 2024), supporting the generality and robustness of our OT-aligned rewards.

J. Hyperparameters Sensitivity Analysis

To further assess robustness, we conduct a stability analysis by sweeping (γ, η) over the range $[0.0, 0.5]$ in 0.1 increments. For each setting, we measure the Video-Text Matching (VTM) accuracy on 10,000 WebVid10M (Bain et al., 2021) video-text pairs using our OT-aligned Semantic Reward with InternVideo2 (Wang et al., 2025a).

We observe that the VTM accuracy varies smoothly across this range, with a standard deviation of only 1.48%, indicating that our method is robust to these weights. This confirms

Table 6. Ablation Study. We analyze the contributions of OT alignment and OT-aligned Rewards to post-training performance. PISCES w/o OT post-trains with pre-trained VLM embeddings, while PISCES w/ $\mathcal{R}_{\text{OT-semantic}}$ and PISCES with $\mathcal{R}_{\text{OT-semantic}}$ assess the impact of OT-aligned Quality and Semantic Rewards, respectively. Full PISCES, which integrates OT alignment and both rewards, achieves the best performance across Total, Quality, and Semantic Scores, demonstrating the effectiveness of structured reward optimization. Bold numbers denote the best results in each category.

Method	OT	$\mathcal{R}_{\text{semantic}}$	$\mathcal{R}_{\text{quality}}$	**Total Score**	**Quality Score**	Subject Consist.	BG Consist.	Temporal Flicker.	Motion Smooth.	Aesthetic Quality	Dynamic Degree	Image Quality
VideoCrafter2	✗	✗	✗	80.44	82.20	96.85	98.22	98.41	97.73	63.13	42.50	67.22
PISCES w/o OT	✗	✓	✓	81.92	83.44	96.99	97.66	98.02	97.16	66.39	52.78	70.50
PISCES w/ $\mathcal{R}_{\text{OT-quality}}$	✓	✗	✓	82.21	**83.77**	97.07	97.58	97.72	97.10	**66.92**	**58.06**	**70.56**
PISCES w/ $\mathcal{R}_{\text{OT-semantic}}$	✓	✓	✗	81.70	82.87	**97.59**	**98.61**	98.06	**97.31**	66.83	40.00	70.19
PISCES	✓	✓	✓	**82.51**	83.73	96.61	97.49	**98.72**	96.80	66.07	**57.50**	70.39

Method	OT	$\mathcal{R}_{\text{semantic}}$	$\mathcal{R}_{\text{quality}}$	**Semantic Score**	Object Class	Multiple Objects	Human Action	Color	Spatial Relation.	Scene	Appear. Style	Temporal Style	Overall Consist.
VideoCrafter2	✗	✗	✗	73.42	92.55	40.66	95.00	**92.92**	35.86	55.29	**25.13**	25.84	28.23
PISCES w/o OT	✗	✓	✓	75.82	95.57	53.06	97.80	90.52	39.51	59.07	24.27	26.03	28.26
PISCES w/ $\mathcal{R}_{\text{OT-quality}}$	✓	✗	✓	75.97	94.84	57.80	**98.00**	90.36	38.51	55.54	24.45	26.37	28.62
PISCES w/ $\mathcal{R}_{\text{OT-semantic}}$	✓	✓	✗	76.99	95.32	59.36	96.60	91.06	**44.97**	**59.93**	24.04	25.81	28.26
PISCES	✓	✓	✓	**77.63**	**98.13**	**66.51**	97.60	92.46	36.07	58.75	23.53	**26.97**	**29.10**

Table 7. **Ablation study on Partial OT and spatio-temporal constraints.** We report Video-Text Matching (VTM) accuracy on 10,000 video-text pairs from WebVid10M. Partial OT with $m = 0.9$ achieves the best alignment, while our spatio-temporal constraints ($\gamma = 0.2$, $\eta = 0.2$) further boost performance. Arrows indicate relative change compared to vanilla cross-attention baseline.

Method	**VTM Acc. (%) ↑**	**Change**
Vanilla cross-attention	81.25	–
Full OT ($m = 1.0$)	86.87	↑5.62
Partial OT ($m = 0.5$)	78.92	↓2.33
Partial OT ($m = 0.9$)	87.54	↑6.29
Partial OT ($m = 0.9$) + spatial only ($\eta = 0.2$)	88.17	↑6.92
Partial OT ($m = 0.9$) + temporal only ($\gamma = 0.2$)	87.98	↑6.73
Partial OT ($m = 0.9$) + spatio-temporal ($\gamma = \eta = 0.1$)	87.06	↑5.81
Partial OT ($m = 0.9$) + spatio-temporal ($\gamma = \eta = 0.2$)	**89.36**	**↑8.11**

Table 8. **Effect of Motion Guidance on post-training VideoCrafter2 and HunyuanVideo**. PISCES significantly outperforms T2V-Turbo-v2 (Li et al., 2025a) in both with and without motion guidance settings.

Models	**VideoCrafter2** (Chen et al., 2024)			**HunyuanVideo** (Kong et al., 2025)		
	Total	**Quality**	**Semantic**	**Total**	**Quality**	**Semantic**
Vanilla	80.44	82.20	73.42	83.24	85.09	75.82
T2V-Turbo-v2 w/o motion (Li et al., 2025a)	81.87 ↑1.43	83.26 ↑1.06	76.30 ↑2.88	84.25 ↑1.01	85.93 ↑0.84	77.52 ↑1.70
PISCES (Direct Backpropagation) w/o motion	**82.51** ↑2.07	**83.73** ↑1.53	**77.63** ↑4.21	**85.05** ↑1.81	**86.84** ↑1.75	**77.89** ↑2.07
T2V-Turbo-v2 w/ motion (Li et al., 2025a)	82.34 ↑1.90	83.93 ↑1.73	75.97 ↑2.55	84.50 ↑1.26	86.32 ↑1.23	77.24 ↑1.42
PISCES (Direct Backpropagation) w/ motion	**82.79** ↑2.35	**84.12** ↑1.92	**77.45** ↑4.03	**85.24** ↑2.00	**87.07** ↑1.98	**77.94** ↑2.12

the stability of our Partial OT formulation with respect to its structured penalty terms. As shown in Figure 8, the variation is minimal and has a negligible impact on performance.

K. Failure Example of Discrete Partial OT

We provide a qualitative example in Figure 9 that illustrates both the sensitivity and limitations of OT alignment. **Left:** The attention map from vanilla cross-attention fails to ground the token *"glasses"*. **Middle:** Our OT plan $\mathbf{P}^\star$ with

Table 9. **Comparison of Evaluators for PISCES on VBench.** Both InternVideo2 and ViCLIP as evaluators yield strong improvements over the Vanilla baseline, confirming PISCES' robustness across feature extractors.

Models	VideoCrafter2 (Chen et al., 2024)			HunyuanVideo (Kong et al., 2025)		
	Total	Quality	Semantic	Total	Quality	Semantic
Vanilla	80.44	82.20	73.42	83.24	85.09	75.82
PISCES w/ InternVideo2 (Wang et al., 2025a)	82.75 ↑2.31	84.05 ↑1.85	77.54 ↑4.12	85.45 ↑2.21	86.73 ↑1.64	80.33 ↑4.51
PISCES w/ ViCLIP (Wang et al., 2024)	82.84 ↑2.40	84.18 ↑1.98	77.49 ↑4.07	85.33 ↑2.09	86.77 ↑1.68	79.58 ↑3.76

Table 10. **OOD Alignment Performance using ViCLIP (Wang et al., 2024).** PISCES achieves the highest cosine similarity, confirming robustness under distribution shifts.

Method	Cosine Similarity ↑
HunyuanVideo (Kong et al., 2025)	0.4128
T2V-Turbo-v2 (Li et al., 2024)	0.4390
VideoReward-DPO (Liu et al., 2025b)	0.4285
PISCES	**0.4517**

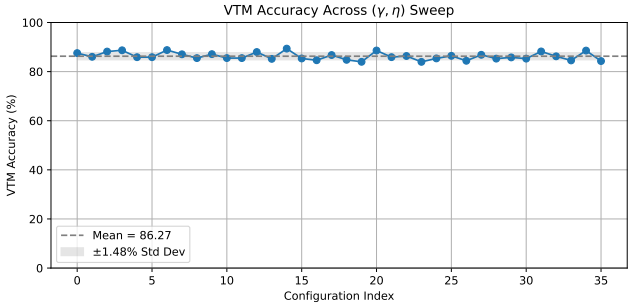

Figure 8. VTM accuracy under (γ, η) sweep. Despite configuration variations, the accuracy remains stable ($\pm 1.48\%$ std).

Figure 9. **Effect of OT mass parameter.** Low OT mass ($m = 0.5$) fails to retain the *"glasses"* token, while higher mass ($m = 0.9$) improves alignment.

partial mass $m = 0.5$ suppresses noise but also removes the valid token *"glasses"*, reproducing misalignment. **Right:** With mass $m = 0.9$, the OT plan retains the *"glasses"* token and aligns it better–but still imperfectly, activating only the right lens.

This reveals a current limitation of the setup: while the OT module improves semantic alignment through structured grounding, overall performance depends on the representational precision of the underlying video–text model, InternVideo2 (Wang et al., 2025a), which was not optimized for fine-grained localization tasks such as segmentation. As a result, even a well-structured transport plan may not fully resolve detailed grounding errors when the base features do not capture sufficient spatial detail. Our approach remains agnostic to the underlying base model, and continued progress in open-source VLMs is likely to further enhance fine-grained grounding performance.

Table 11. **Adaptive Reward Fusion.** We compare equal and adaptive weighting of reward signals in direct backprop. Adaptive weighting using Group Relative Reward improves performance.

Method	Total	Quality	Semantic
HunyuanVideo	83.24	85.09	75.82
PISCES-direct (equal)	85.05 ↑1.81	86.84 ↑1.75	77.89 ↑2.07
PISCES-direct (adaptive)	**85.16** ↑1.92	**86.92** ↑1.83	**78.11** ↑2.29

L. Human Preference Details

To assess inter-rater reliability, we computed Fleiss' Kappa over the collected human preference judgments. Across the three evaluation axes (visual quality, motion quality, and semantic alignment), we obtained an average Fleiss' Kappa score of 0.72, indicating substantial agreement among raters. All raters were blinded to method identity to prevent bias.

We analyzed category-level bias by classifying 400 prompts into 290 motion-heavy (e.g., running, jumping) and 110 static (e.g., portraits, scenic shots). The Pearson correlation between category type and preference for our method is 0.038, indicating no significant bias–our method performs consistently across both motion-heavy and static prompts.

M. Reward Fusion

In our default direct backpropagation setup, we equally weight the consistency loss and reward signals. Moreover, we explore an adaptive weighting strategy using the Group-Relative Reward formulation from GRPO (DeepSeek-AI et al., 2026; Liu et al., 2025a), normalizing reward values across generations for the same prompt (subtracting the mean and dividing by standard deviation). We apply this normalization in the direct backpropagation setting without using RL. Results on HunyuanVideo (Kong et al., 2025) (Table 11) indicate that adaptive weighting via Group-Relative Reward offers a consistent improvement over equal weighting, particularly in semantic alignment and overall score, confirming its potential as a robust reward fusion strategy.

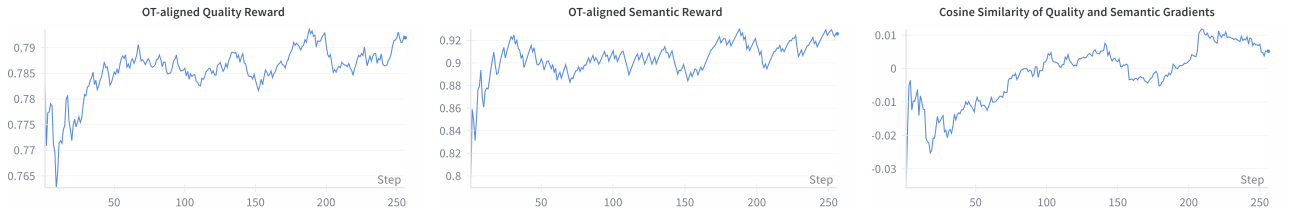

Figure 10. Reward and gradient trends in post-training.

N. Dual Rewards Interaction

To assess the interaction between the OT-aligned quality and semantic rewards during training, we plot their reward trajectories and the cosine similarity of their gradients over time in Figure 10. We observe that both rewards improve steadily across training steps. At step 256 (final), the semantic reward reaches a high value of 0.9268, and the quality reward converges to 0.7814, suggesting the model successfully learns to optimize for both objectives. The cosine similarity of the gradients between the two rewards remains near zero throughout training (final value 0.0074), with small fluctuations in both directions. This implies that the two objectives provide largely orthogonal supervision signals and do not interfere with each other. There is no sign of reward conflict or mode collapse. Together, these findings indicate that our dual reward formulation, one distributional and one token-level, are both stable. It effectively supports the joint optimization of visual quality and semantic alignment in text-to-video generation. This stability enables consistent improvements across both dimensions without sacrificing one for the other.

O. Reward Hacking and Mitigation Strategy

We found that the Consistency Distillation (CD) (Song et al., 2023; Wang et al., 2023b; Lu & Song, 2025) loss helps mitigate reward hacking and stabilize training. Intuitively, this loss anchors the student model (updated via reward supervision) to the teacher model's original distribution, serving both as a regularizer and as an efficient training mechanism by enabling fewer denoising steps. We evaluate the impact of CD loss in Table 12. Without CD loss, the semantic reward is optimized more aggressively, leading to a noticeable drop in visual quality (from 86.84 to 86.51) and total score (from 85.05 to 84.80), confirming that the model may over-optimize the reward signal at the expense of generation fidelity. By retaining CD loss, we balance the original model performance and reward-driven improvements.

P. Training Cost and Efficiency Analysis

To train the OT map (Korotin et al., 2023), we use video-text pairs with 8-frame clips, extracted using frozen Intern-Video2 (Wang et al., 2025a), and train on a single A100 GPU for one day, equivalent to 24 A100 GPU-hours. In comparison, annotation-based pipelines such as VideoReward-DPO (Liu et al., 2025b) require 72 A800 GPU-hours to train their reward models.

Table 12. **Impact of CD Loss on Reward Stability and Quality.** CD loss serves as a regularizer that prevents overfitting to reward signals and stabilizes training.

Method	Total Score	Quality	Semantic
HunyuanVideo (Kong et al., 2025)	83.24	85.09	75.82
PISCES w/o CD Loss	84.80 ↑1.56	86.51 ↑1.42	**77.95** ↑2.13
PISCES w/ CD Loss	**85.05** ↑1.81	**86.84** ↑1.75	77.89 ↑2.07

The total wall-clock cost of PISCES post-training is 29.78 GPU-hours on 8×A100s, slightly higher than the 26.52 GPU-hours required by T2V-Turbo-v2 (Li et al., 2025a) (training time only, without evaluation). While PISCES incurs a marginal increase in training time, this is a one-time cost and remains negligible compared to the massive pre-training costs of T2V models-665,000 GPU-hours for Seaweed-7B (Seaweed et al., 2025) and even more for HunyuanVideo-13B (Kong et al., 2025). Thus, PISCES provides a lightweight and effective reward alignment mechanism with minimal computational overhead.

Finally, inference incurs no additional cost from the reward models. In fact, for GRPO-tuned models, we reduce the denoising steps from 50 to 16 through consistency distillation, resulting in approximately 3× faster inference.