Dual-IPO

Dual-Iterative Preference Optimization for Text-to-Video Generation

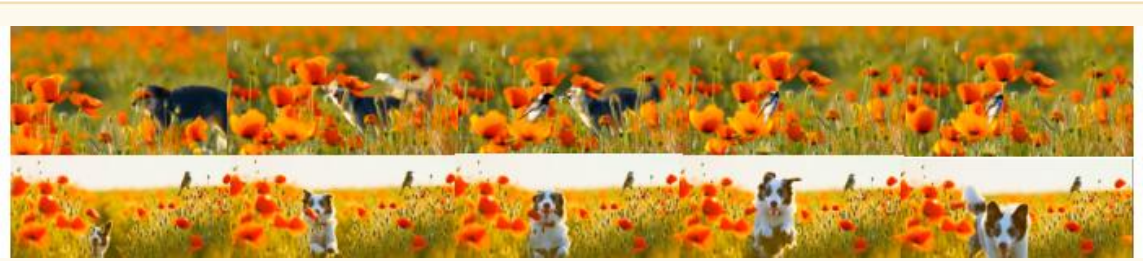
Iteratively improve both the critic and the generator
for stronger human alignment in text-to-video generation.

Xiaomeng Yang, Mengping Yang, Jia Gong, Luozheng Qin, Zhiyu Tan, Hao Li

Fudan University • Shanghai Academy of AI for Science

Paper Overview

Representative qualitative cases from the paper



Key message: reward signals should evolve with the generator rather than stay fixed.

Motivation: Why Static Preference Learning Falls Short

The reward signal should evolve with the generator as failure modes change during post-training.

Challenges in current T2V post-training

1 T2V models still miss user preferences

Prompt mismatch, motion artifacts, and weak aesthetics remain common even for strong diffusion transformers.

2 Human preference data is expensive to scale

Large manually annotated datasets are laborious to build and quickly become the bottleneck for alignment.

3 Fixed rewards and static datasets go stale

As the generator improves, old rewards no longer capture the model's latest failure cases or subtle artifacts.

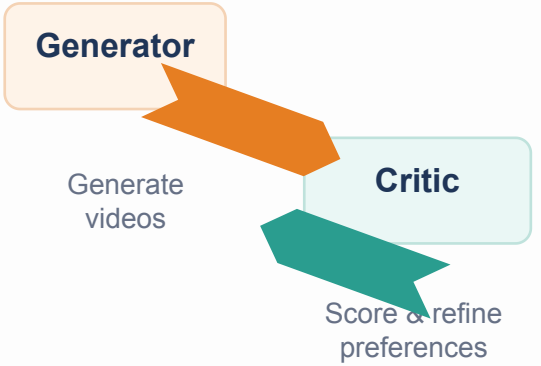
From static alignment to dual iteration

Static pipeline

Prompt
→ Generator
→ Fixed reward
→ One-shot DPO/KTO

Reward model and preference data stay frozen.

Dual-IPO insight



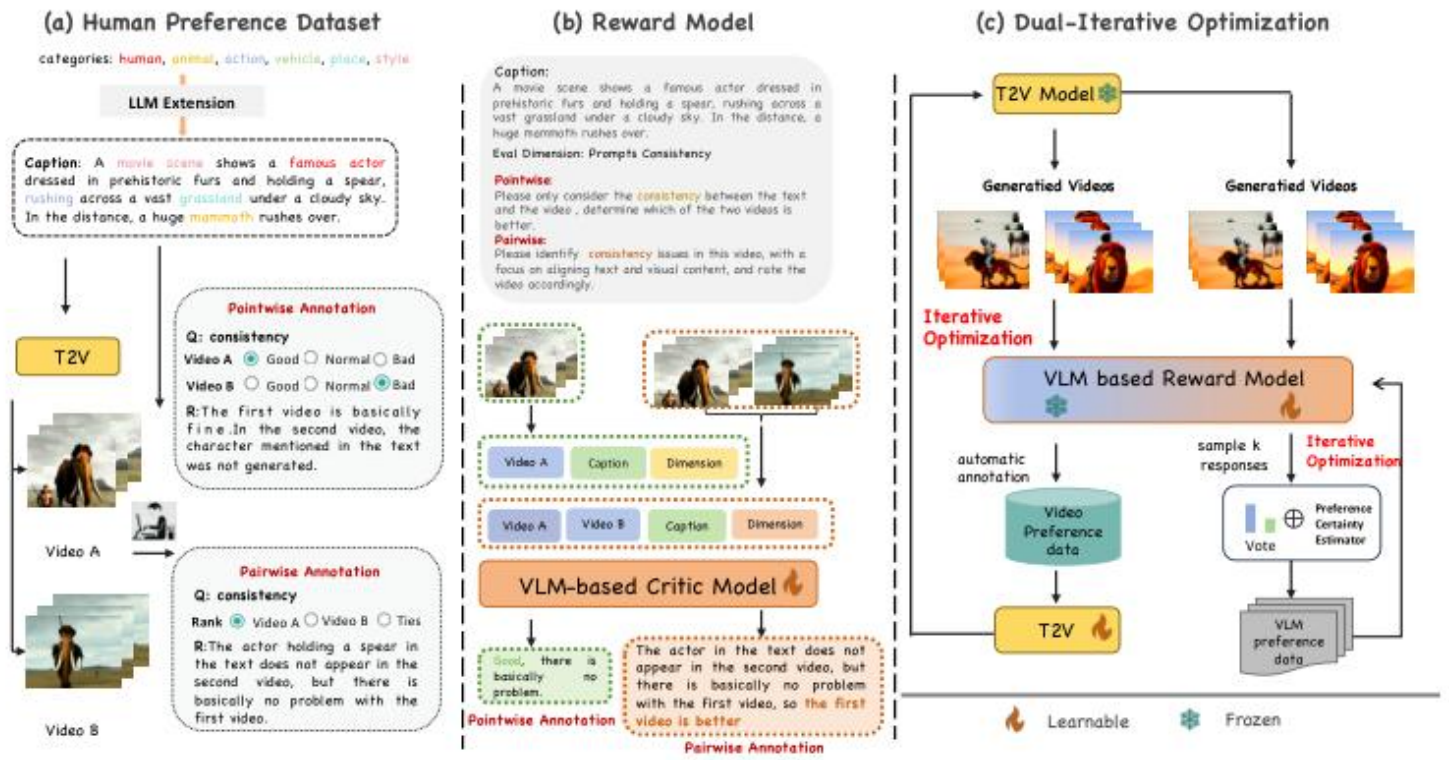
- Train a critic from a small seed of human preferences.
- Use the critic to annotate new generations automatically.
- Alternate between critic refinement and generator alignment.

Dynamic reward feedback rebuilds the preference dataset each round.

Need a framework where both the supervisor and the generator improve together.

Dual-IPO Framework Overview

A small human preference seed trains a VLM-based critic, which then co-evolves with the T2V generator across multiple rounds.



1. Seed human preferences

Collect a small set of pairwise and pointwise labels across consistency, faithfulness, and motion.

2. SRPO builds a stronger critic

Use CoT-guided reasoning, self-consistency voting, and confidence filtering to create robust pseudo-labels.

3. Critic and generator co-evolve

Re-score new generations, rebuild preference data, and refine the T2V model in repeated rounds.

Self-Refined Preference Optimization (SRPO)

SRPO improves the critic with structured reasoning, self-consistency, and confidence-aware filtering.

Critic-side refinement

The reward model is not static: it is updated with pseudo-labels generated by the latest critic itself.

Goal

Produce stronger and more reliable reward signals with minimal manual annotation.

1

CoT-guided annotation

Warm-start the critic with Chain-of-Thought supervision on a small seed set, using structured reasoning to judge preference dimensions.

2

Self-consistency voting

Sample multiple reasoning paths and vote across answers, reducing noise from any single inference trace.

3

Preference Certainty Estimator

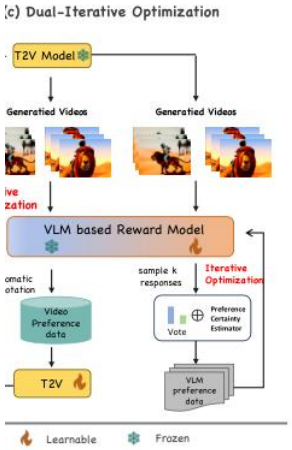
Keep only high-confidence preferences so the critic focuses on reliable supervision under distribution shift.

Iterative Generator Alignment

Use the updated critic to produce fresh preference supervision for the latest generator.

Dynamic alignment loop

Dual-IPO keeps refreshing supervision from the latest critic and the latest generator outputs.



Pairwise: Diffusion-DPO **Pointwise: Diffusion-KTO**

The framework activates either path depending on the available preference format.

What changes in each round?

Step 1

Generate candidates

For each prompt, the T2V model samples multiple candidate videos.

Step 2

Score and rank

The critic assigns pairwise or pointwise preferences to the new generations.

Step 3

Optimize and monitor

The generator is updated with DPO/KTO, while VBench and loss dynamics are monitored.

Step 4

Refresh the critic when needed

If reward misalignment or degradation appears, SRPO updates the critic and rebuilds the dataset.

Outcome: the reward signal adapts as the generator's failure modes become subtler.

Experimental Setup

Dual-IPO is evaluated across different T2V backbones with both automatic and human metrics.

Critic model

VILA 13B/40B reward model
16 frames per video
CoT warm-start + 5 epochs of training
~20K pseudo-labels generated per iteration

Generator alignment

8 candidates per prompt
Top-1 vs. bottom-1 form preference pairs
Batch size 512 across 128 GPUs
One round takes about two weeks

Data

~1M synthetic captions from structured elements + LLM expansion
VidGen-1M subset added for real-video regularization
~100K preference pairs per round

Evaluation

VBench overall + per-dimension scores
Human evaluation on consistency, motion, and faithfulness
Baselines: CogVideoX-2B/5B and Wan-1.3B

Evaluation focus

Does Dual-IPO improve the reward model itself, deliver steady gains over multiple rounds, and generalize across different T2V architectures and scales?

Reward model quality

Compare against generic open-source video reward models with human preference accuracy.

Iterative improvement

Track VBench across repeated IPO rounds and compare against static one-round optimization.

Cross-model generalization

Apply the same framework to CogVideoX-2B, CogVideoX-5B, and Wan-1.3B.

Quantitative Results

Dual-IPO improves multiple generators and achieves strong performance against current T2V baselines.

Comparison with strong T2V systems

Model	Total	Quality	Semantic
Vidu Q1	87.41	87.28	87.94
Open-Sora-2.0	84.34	85.40	80.12
Sora	84.28	85.51	79.35
CausVid	84.27	85.65	78.75
Wan2.1	86.22	86.67	84.44
CausVid	83.88	85.21	78.57
HunyuanVideo	83.24	85.09	75.82
CogVideoX-5B	82.01	82.72	79.17
+Ours	**86.57**	**87.00**	**84.84**
CogVideoX-2B	80.91	82.18	75.83
+Ours	**82.74**	**83.92**	**78.00**
Wan-1.3B	84.26	85.30	80.09
+Ours	**88.32**	**87.83**	**90.25**

Key findings

Customized critic matters

The paper reports that the Dual-IPO critic outperforms generic video reward models and produces better downstream alignment.

Generalizes across generators

The same framework improves CogVideoX-2B, CogVideoX-5B, and Wan-1.3B instead of overfitting to a single backbone.

Efficiency at smaller scale

After Dual-IPO, a 2B CogVideoX model can outperform a baseline 5B model, showing strong post-training efficiency.

Top-end result

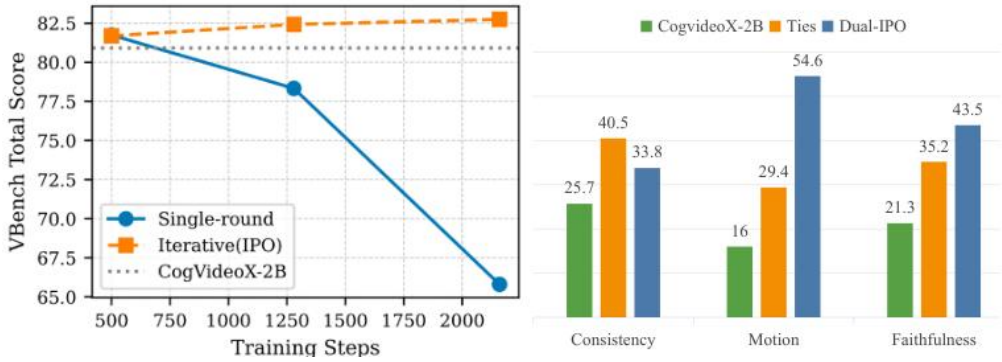
Wan-1.3B + Ours achieves the strongest overall VBench result in the comparison table.

Overall takeaway: Dual-IPO brings robust gains across model sizes and architectures, not just on a single benchmark setting.

Why Iteration Matters

Dual-IPO keeps improving while static one-round optimization degrades under stale rewards.

Progression and human evaluation



Static single-round DPO improves briefly, then degrades; Dual-IPO continues to climb and is preferred in human evaluation.

Iteration ablation

Models	Total Score.	Quality Score.	Semantic Score.
CogVideoX-2B	80.91	82.18	75.83
CogVideoX-2B-IPO$_1$	81.69	82.87	77.01
CogVideoX-2B-IPO$_2$	82.42	83.53	77.97
CogVideoX-2B-IPO$_3$	82.74	83.92	78.00

Performance improves steadily across IPO rounds, validating repeated critic-generator co-optimization.

SRPO ablation

Setting	Critic Update	SRPO loss	PCE	VBench Score
Baseline	✗	—	—	82.74
No update	✗	—	—	82.33
SRPO (full)	✓	✓	✓	**82.91**
SRPO loss	✓	✗	✓	82.83
PCE	✓	✓	✗	82.69

- Full SRPO performs best during later iterations.
- Removing critic updates hurts the final score.
- SRPO loss and PCE both contribute to the gain.

Qualitative Results

Visual comparisons show stronger prompt following, smoother motion, and better subject consistency after Dual-IPO.

Representative comparisons from the paper

What changes visually?

Prompt following
The generated content better matches the prompt intent and its detailed scene description.

Subject consistency
Subjects remain more stable across frames with fewer structural errors or identity drift.

Motion smoothness
Temporal transitions are more natural and less artifact-prone during the sequence.

Aesthetic quality
Outputs look cleaner and more visually coherent, with stronger overall realism.

Takeaways

Dual-IPO offers a simple and effective recipe for post-training text-to-video models.

What Dual-IPO does

- Jointly optimize the critic and the generator instead of freezing either side.
- Alternate between reward-model refinement and T2V alignment in multiple rounds.
- Support both pairwise DPO and pointwise KTO under one framework.

Why it works

- Warm-start the critic with CoT-guided reasoning on a small human seed set.
- Use self-consistency voting and PCE filtering to improve pseudo-label reliability.
- Keep reward supervision synchronized with the generator's evolving error patterns.

What it achieves

- Consistent gains on CogVideoX-2B, CogVideoX-5B, and Wan-1.3B.
- A post-trained 2B model can outperform a baseline 5B variant.
- Qualitative and human studies show better alignment, motion, and realism.

Closing message: Iteratively improving both the supervisor and the generator is a powerful way to align text-to-video models with human preferences.