

PhysHPO: Hierarchical Fine-grained Preference Optimization for Physically Plausible Video Generation

Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, Ser-Nam Lim

The Hong Kong University of Science and Technology,

University of Central Florida

Project Page: <https://haroldchen19.github.io/PhysHPO-Page/>

Presenter: Harold Haodong Chen

11/2025

Motivation: Beautiful But Physically Wrong

- ***Problem:*** Current T2V models produce visually appealing but physically implausible videos (e.g., momentum violations, fluid inconsistencies)
- ***Why it matters:*** world model, simulation, scientific and educational content, *etc.*
- ***Limitations of prior approaches:***
 - ❑ Test-time LLM-refinement (e.g., PhyT2V): slow, bounded by base model;
 - ❑ SFT (e.g., WISA, SynVideo): costly data creation, fixed signals;
 - ❑ DPO for video: mostly instance-level, too coarse.

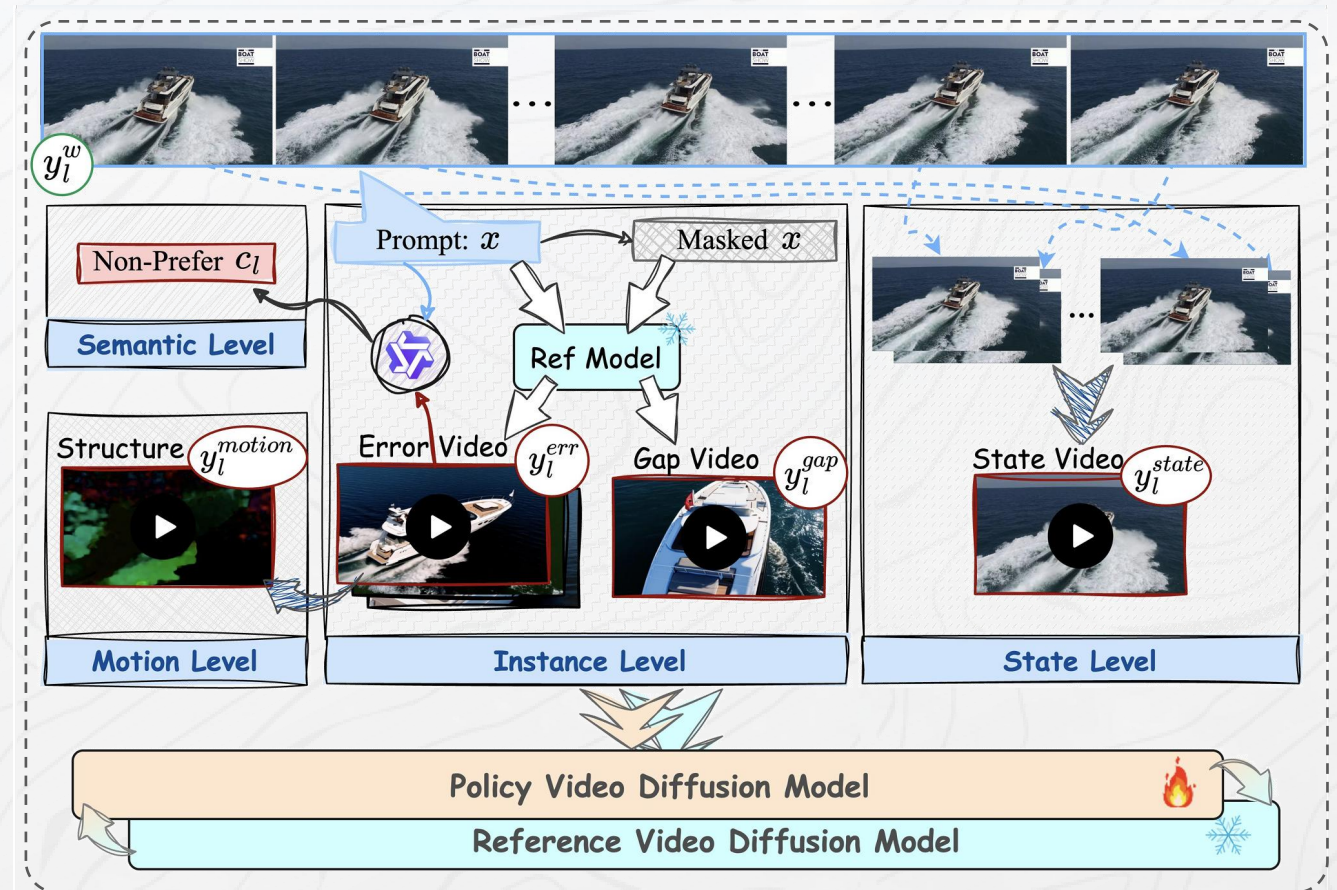
Core Idea in One Picture

- ***PhysHPO = Hierarchical Cross-Modal DPO with 4 levels:***

- Instance-level overall preference
- State-level boundary preference
- Motion-level dynamic preference
- Semantic-level consistency preference

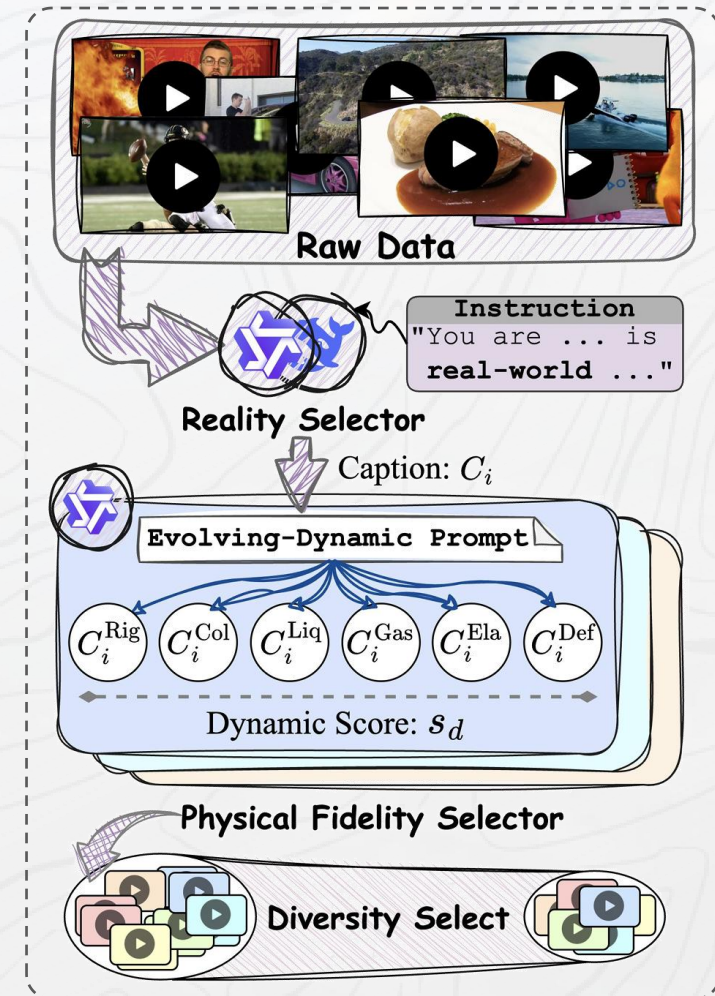
- ***Data Selection → Real-world videos are the best physics teachers:***

- Reality
- Physics-rich
- Diverse



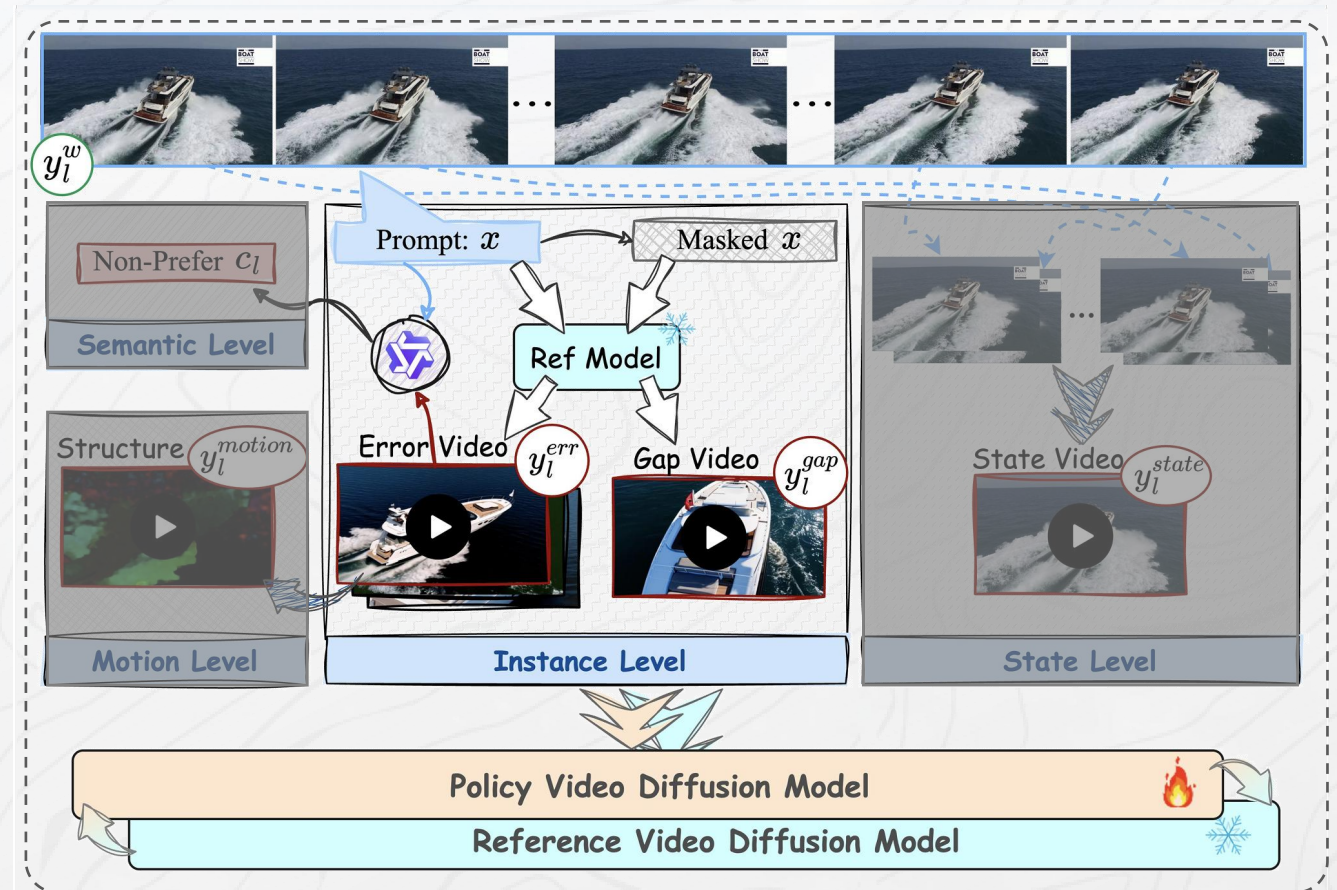
Selecting “Good Data” Instead of Building New

- *Starting from a large high-quality dataset (e.g., OpenVid)*
- *Step 1: Reality filter*
 - ❑ VLM-based double-check → near 100% accuracy
- *Step 2: Physical fidelity scoring:*
 - ❑ 17 phenomena across Dynamics, Thermodynamics, Optics;
 - ❑ Evolving prompts to expand each caption across dimensions;
- *Step 3: Diversity selection:*
 - ❑ Embedding-based cosine diversity with a high threshold;
 - ❑ Final: compact, high-fidelity, diverse subset



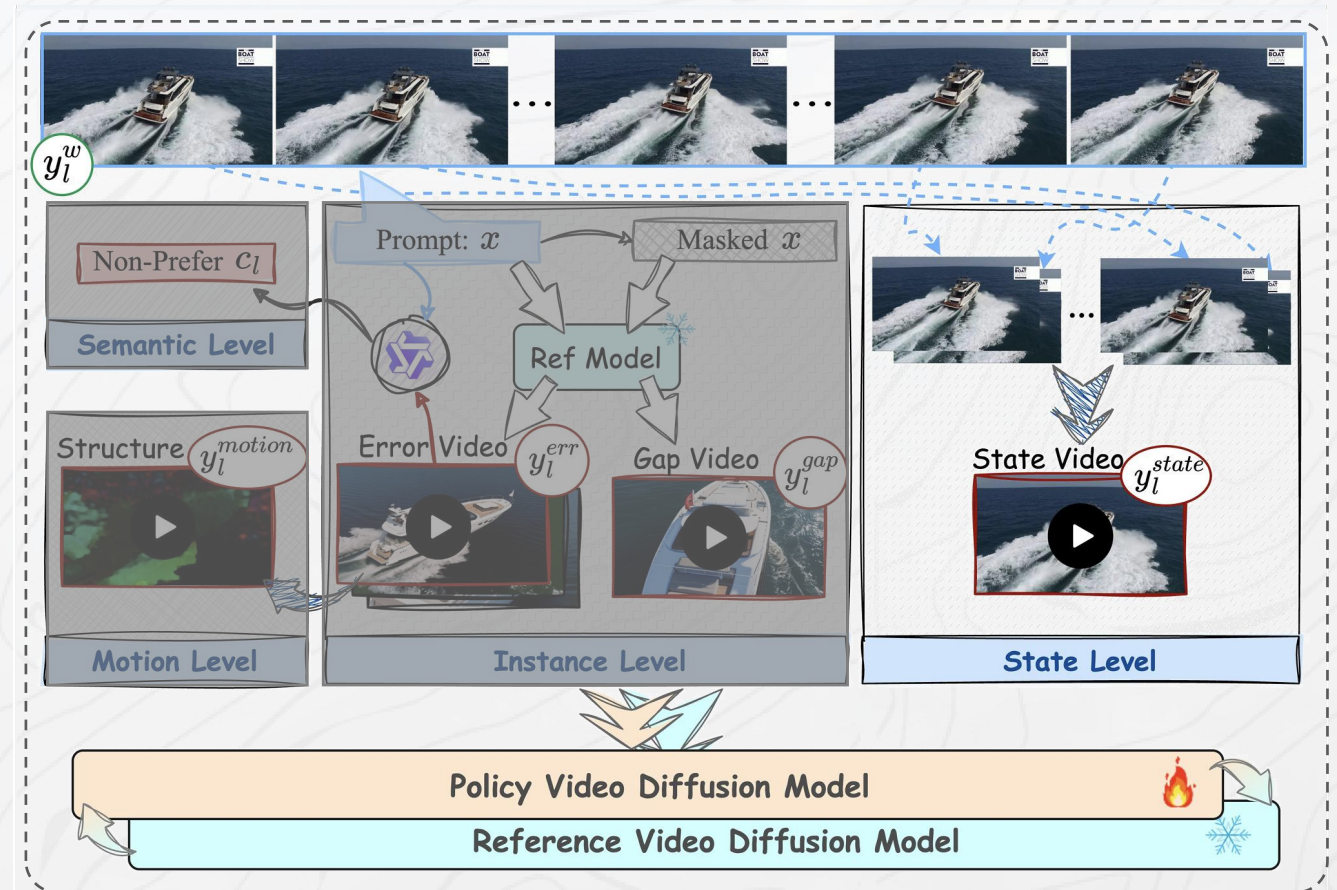
Instance-Level Overall Preference

- *Preferred Video: from selected real videos*
- *Two types of non-preferred video per prompt:*
 - **Error Video** y_l^{err} : same semantics, most visually similar but with physical imperfections (from base model);
 - **Gap Video** y_l^{gap} : semantic gaps via prompt masking during generation.



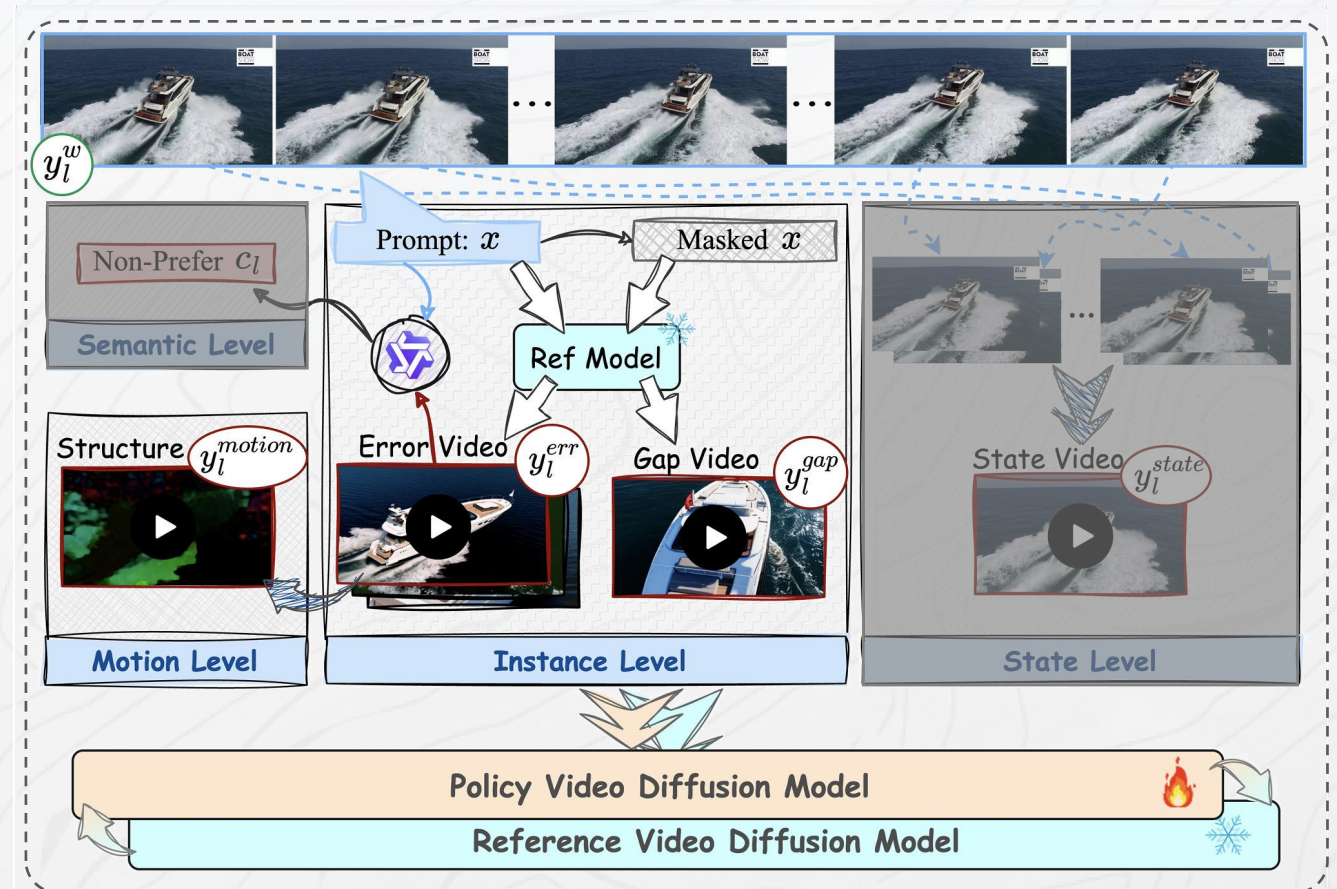
State-Level Boundary Prefernece

- *Replace first/last frames (e.g., $N=2$) to create State Video y_l^{state}*
- *Optimize preference to keep physically consistent start/end states*



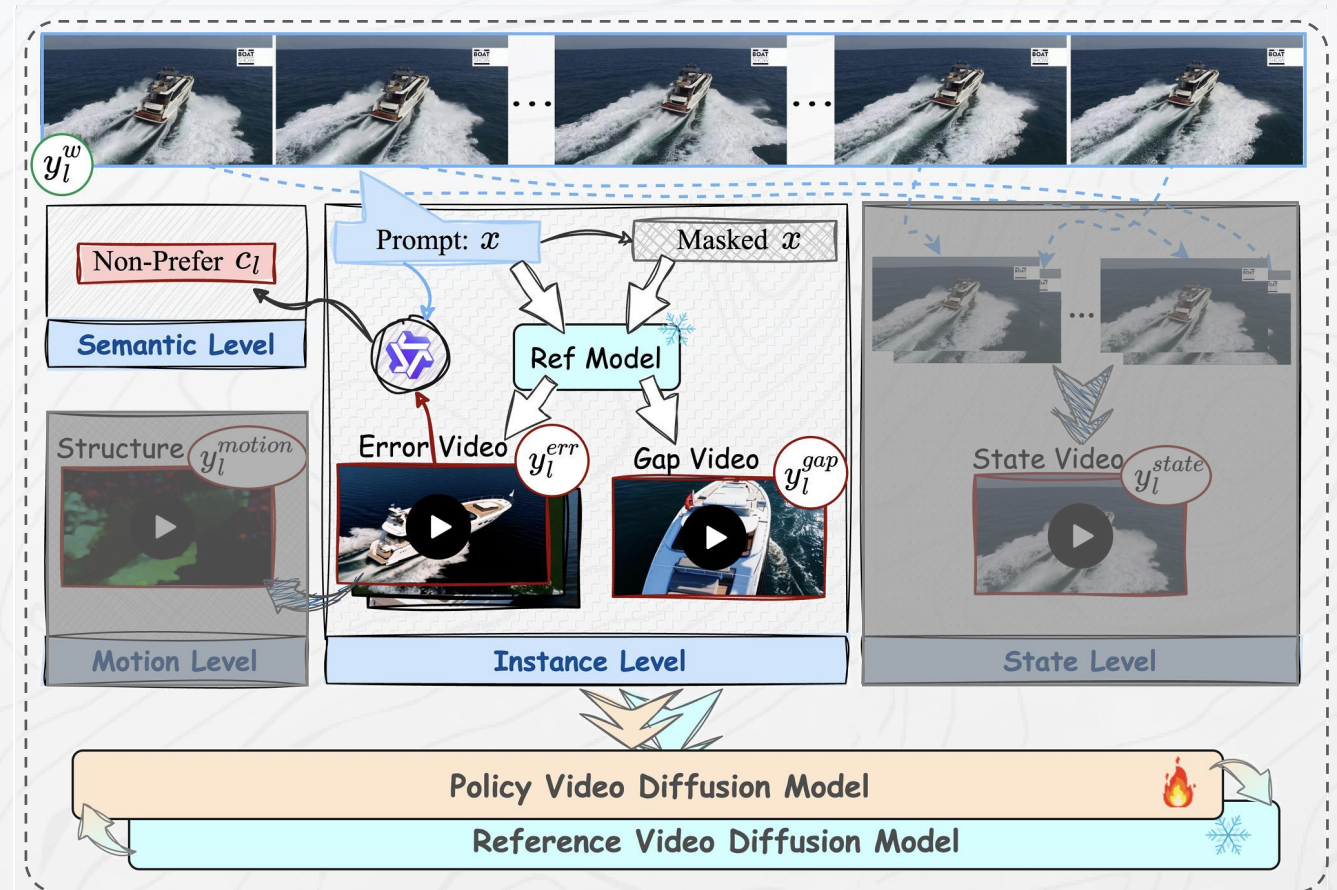
Motion-Level Dynamic Prefernece

- *Extract structural information y_l^{motion} (e.g., trajectories/flow/structure cues) from Preferred Video and Error Video*
- *Optimize preference over motion representations (not just pixels)*



Semantic-Level Consistency Preference

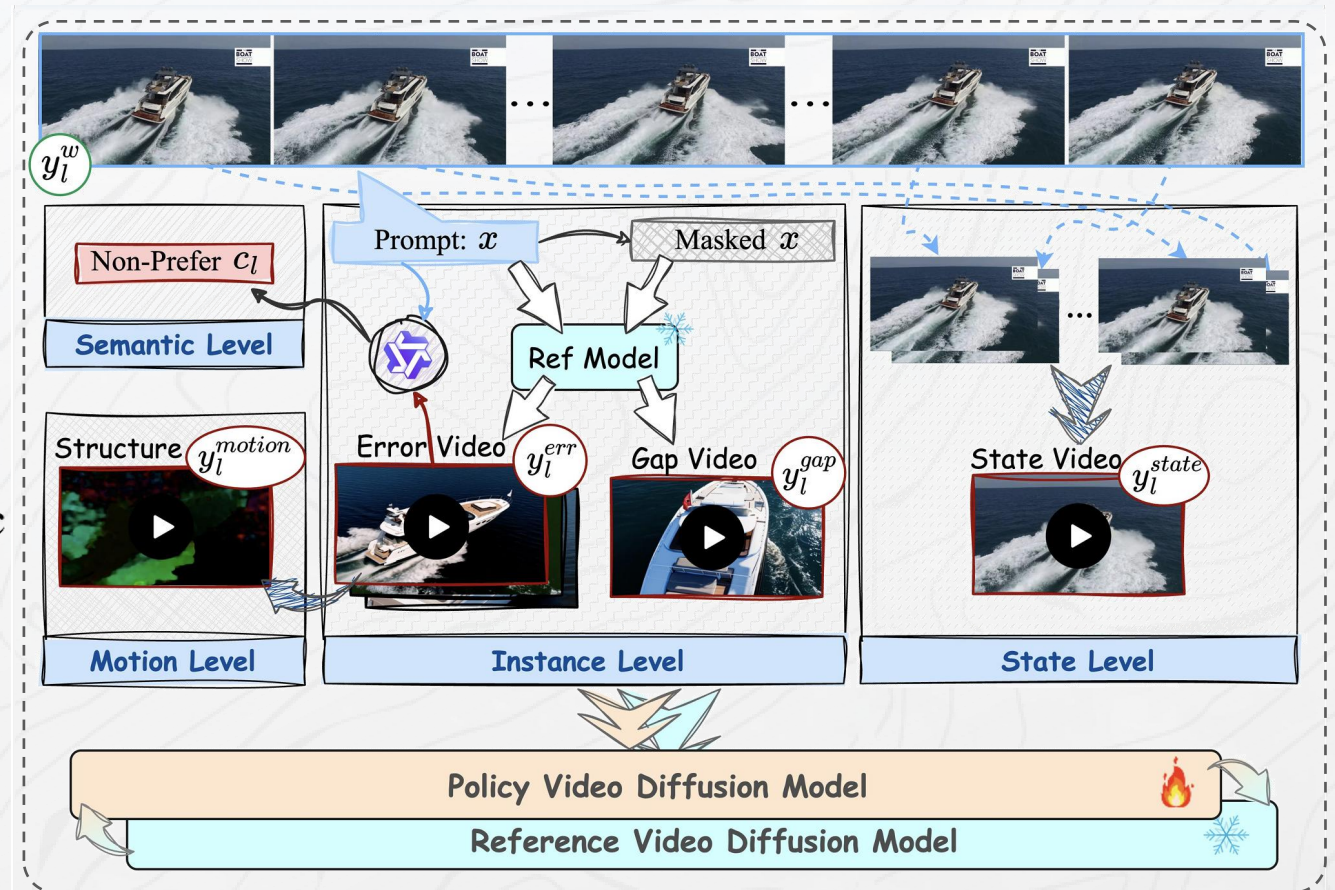
- *Treat the original prompt as Preferred Caption c_w*
- *Use a VLM to minimally edit Preferred Caption to match Error Video's inconsistencies $\rightarrow$ Non-Preferred Caption c_l*
- *Preference compares $p_\theta(y_w|c_w)$ vs. $p_\theta(y_l|c_l)$*



Hierarchical Cross-Modal Preferred Optimization

- **Total Loss:**

$$\mathcal{L}_{\text{PhysHPO}} = \mathcal{L}_{\text{Instance}} + \lambda \mathcal{L}_{\text{State}} + \rho \mathcal{L}_{\text{Motion}} + \mu \mathcal{L}_{\text{Semantic}}$$



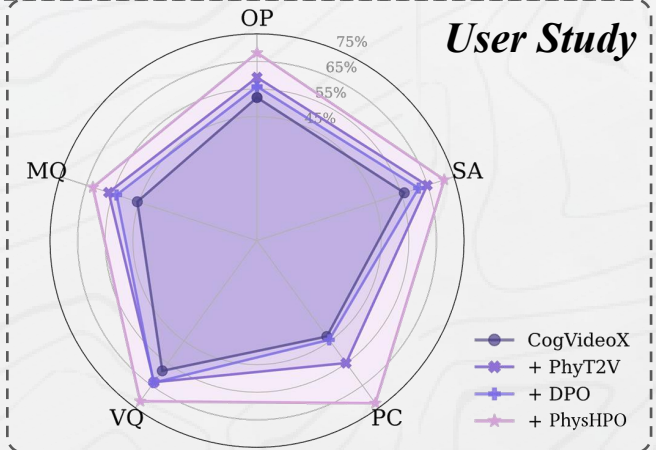
Evaluation on VideoPhy, PhyGenBench, and VBench

Method	VideoPhy [6]				PhyGenBench [54]					VBench [34]		
	SS↑	SF↑	FF↑	Over.↑	Mechanics↑	Optics↑	Thermal↑	Materials↑	Overall↑	Total↑	Quality↑	Semantic↑
CogVideoX-2B [90]	12.7	21.9	25.4	18.6	0.38	0.43	0.34	0.39	0.39	81.6	82.5	77.8
+ PhyT2V [85]	14.1	19.9	28.6	18.9	0.45	0.48	0.34	0.50	0.45	82.0	82.9	78.4
+ Vanilla DPO [72]	15.5	19.2	28.6	19.2	0.43	0.50	0.40	0.50	0.46	82.2	83.1	79.0
+ PhysHPO (Ours)	20.4	24.7	42.9	25.9	0.50	0.56	0.47	0.58	0.53	82.5	83.3	79.3
CogVideoX-5B [90]	24.4	53.1	43.6	39.6	0.39	0.55	0.40	0.42	0.45	81.9	83.1	77.3
+ PhyT2V [85]	25.4	48.6	55.4	40.1	0.45	0.55	0.43	0.53	0.50	82.3	83.3	78.3
+ Vanilla DPO [72]	28.2	50.0	51.8	41.3	0.48	0.60	0.47	0.58	0.54	82.4	83.3	78.7
+ PhysHPO (Ours)	32.4	54.1	58.9	45.9	0.55	0.68	0.50	0.65	0.61	82.8	83.7	79.3

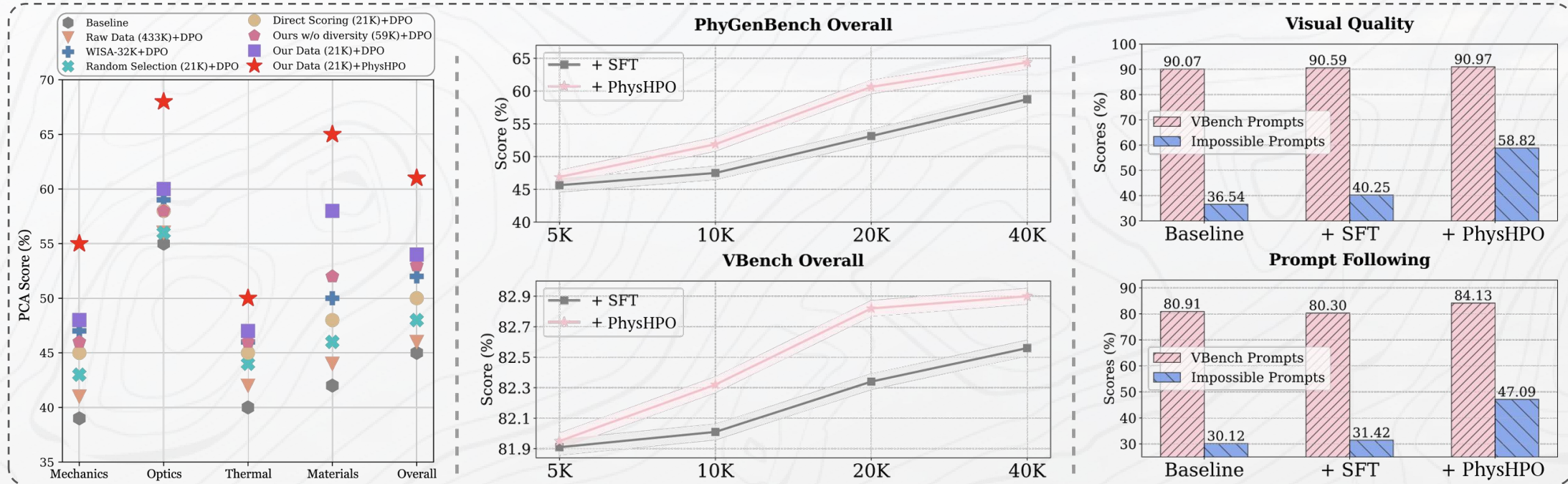
- *PhysHPO demonstrates superior performance in enhancing both physical fidelity and general quality.*
- *PhysHPO effectively align video generation with human preference.*



Key Findings



Analysis of Data Selection and PhysHPO vs. SFT



- Superiority of our data selection strategy & Importance of data diversity.
- PhysHPO is a data-efficient video generation enhancer.
- PhysHPO enables creative generalization beyond physical laws.



Thanks!

Qualitative Results in: <https://haroldchen19.github.io/PhysHPO-Page/>

Presenter: Harold Haodong Chen

11/2025