

# From Video Generation to World Models

Learning to predict, plan, and act

---

**Harry Yang**

Assistant Professor, HKUST · DNA Lab  
UNIST · October 2026

PROMPT

*“A sunflower grows in front of the house.”*



time →

**Sharp. Stable. Beautiful.**  
**And no sunflower.**

Output of FramePack, a strong open image-to-video model.

# Video models have learned to render. Next, they must learn to predict



Make-A-Video samples: “a teddy bear painting a portrait”, “a knight riding a horse”, “a panda on a swing”

- **2022 TATS** ECCV 2022  
Long videos with a time-agnostic VQGAN and time-sensitive transformer  
→ [code & samples](#)
- **2023 Make-A-Video** ICLR 2023 • Meta  
Text-to-video without paired text-video data  
→ [makeavideo.studio](#)
- **2025 VideoGen-of-Thought** NeurIPS 2025 workshop oral  
Multi-shot stories from a single sentence  
→ [demo video](#)
- **2026 This talk** ICML • NeurIPS • ECCV  
From rendering to predicting, planning and acting

- **Making a video look convincing is no longer the hard part. Making it useful for a decision is.**

# A world model is a video generator you can make decisions with

## WORLD MODEL

$$\hat{s}_{t+1} \sim p_{\theta}(\cdot \mid s_t, a_t)$$

**s** state: the frames we observe

**a** action: an instruction, a robot command

**$\hat{s}$**  the future we imagine before acting

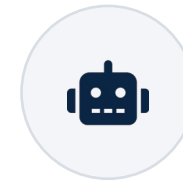
*A video generator conditioned on actions is a natural candidate.*



**Observe**  
current frames



**Imagine**  
futures for each  
candidate action



**Choose & act**  
take the best action



**Repeat**  
with new observations

The imagined future is only useful if it is...



**Controllable**



**Physical**



**Fast**



**Goal-directed**

- Four requirements for an imagination you can plan with.

# Four requirements, four papers



01

## Controllable

Does the prediction change when we ask it to?

---

**AlignVid** ICML 2026  
Training-free attention scaling



02

## Physical

Does the predicted event obey physics?

---

**PhysHPO** NeurIPS 2025  
+ SAGE-GRPO: stable RL for video



03

## Fast

Can we predict in one step?

---

**DMDR** ECCV 2026  
+ Z-Image-Turbo · 4-bit video attention



04

## Goal-directed

Given only a goal, does it know what must change?

---

**SolveEdit** 2026  
Benchmark + test-time planner

● At the end, the four pieces go back into one loop.

# 01

## Controllable

If we change the action, does the prediction change?

### AlignVid

ICML 2026 · Yexin Liu et al.

Project page [law1223.github.io/AlignVid](https://law1223.github.io/AlignVid) Code [github.com/LAW1223/AlignVid](https://github.com/LAW1223/AlignVid)

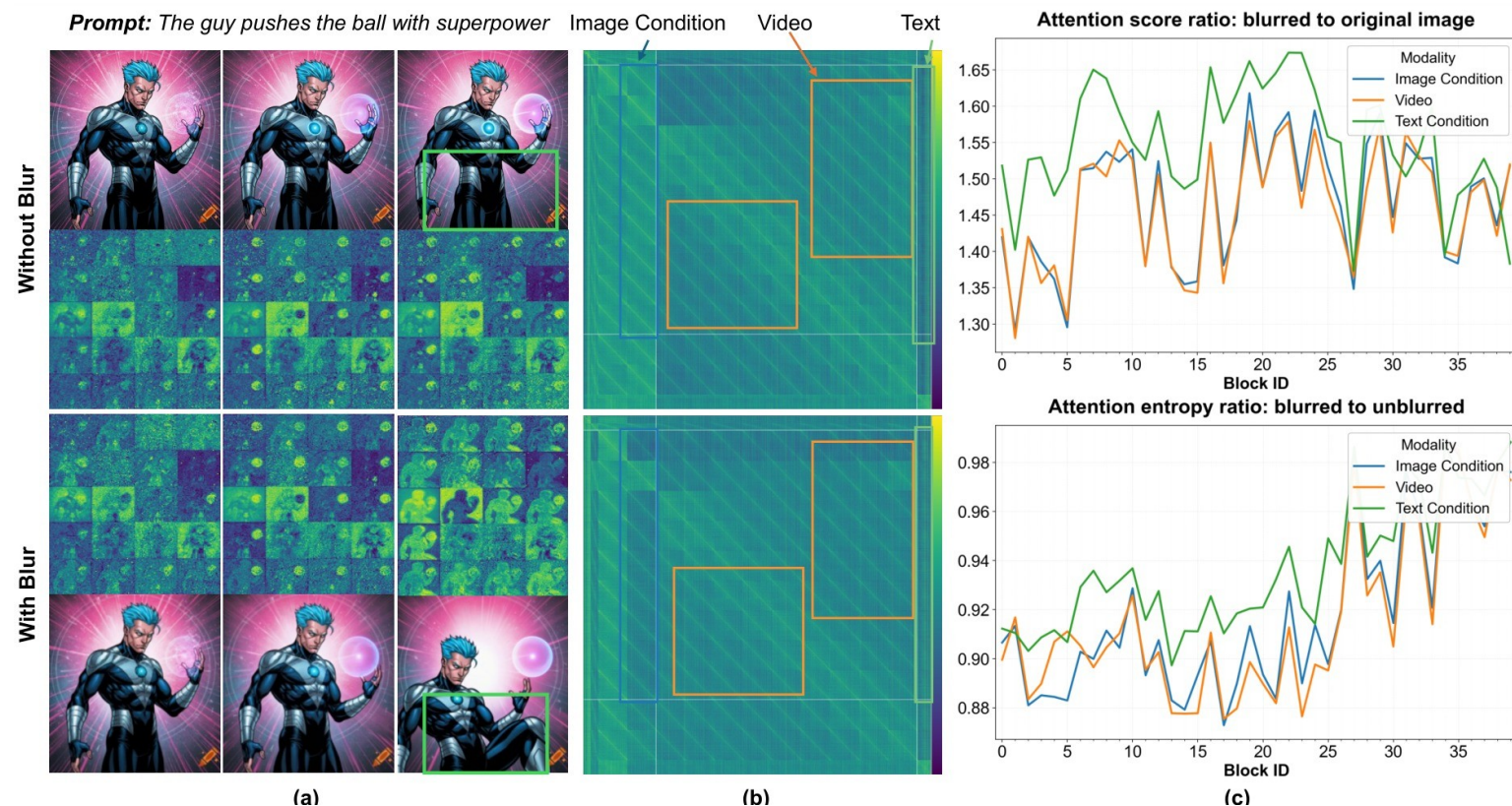


# Same model, same weights. Now the sunflower grows



- The capability was there. The model just wasn't listening.

# A strange clue: blurring the input image helps



## Blur the image

and the requested action happens.

## Attention sharpens

entropy ratio below 1 across all 30 examples.

## So: sharpen attention directly

and keep the image detail.

- Fine image details were competing with the instruction.

# Sharpen attention, but only where and when it matters

## WHAT · Attention Scaling Modulation

$$p_j(\alpha) = \frac{e^{\alpha z_j}}{\sum_k e^{\alpha z_k}}$$
$$\frac{dH}{d\alpha} = -\alpha \operatorname{Var}_{p(\alpha)}(z) \leq 0$$

**Scale logits by  $\alpha > 1$**  before softmax.

**Entropy falls monotonically** in  $\alpha$ : scaling always sharpens.

## WHERE & WHEN · Guidance Scheduling

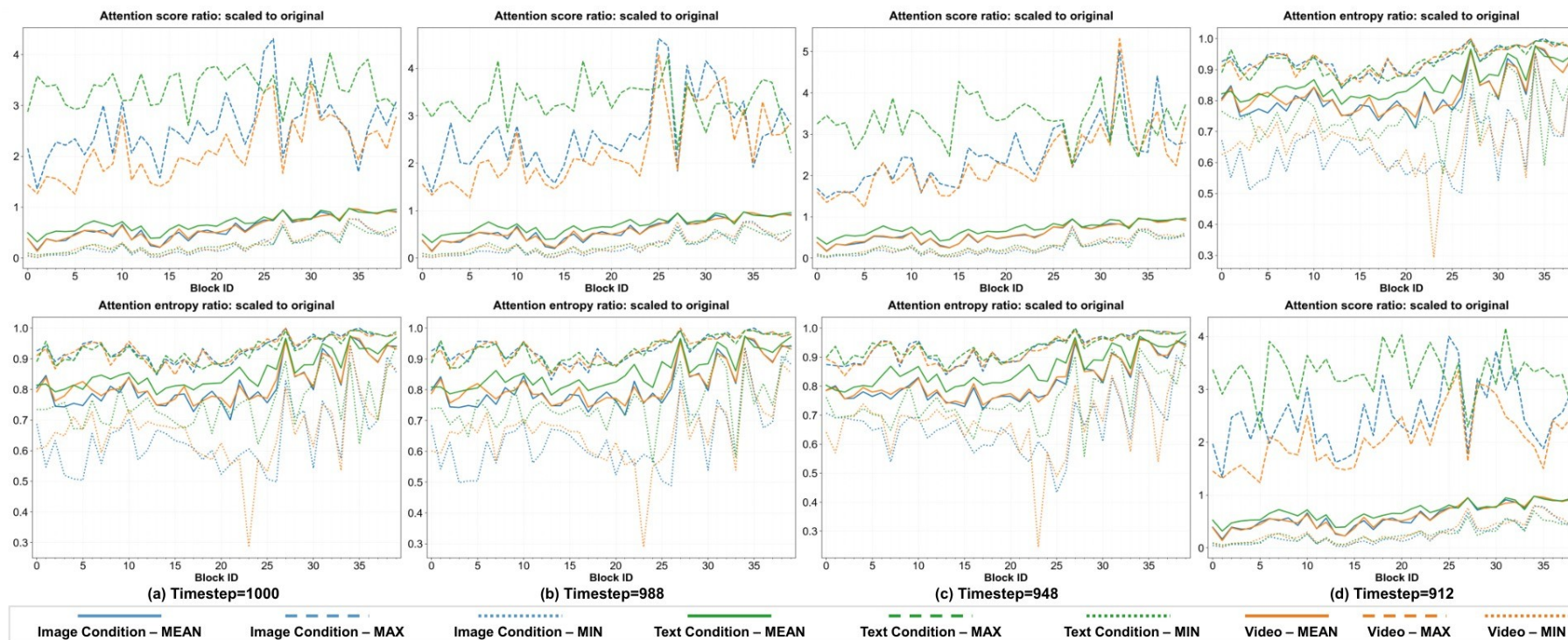
$$g_{\ell,t} = 1 + b_{\ell} m_t (\gamma - 1)$$
$$A_{\ell,t} = \operatorname{softmax} \left( \frac{Q(g_{\ell,t} K)^{\top}}{\sqrt{d}} \right) V$$

**Layers** whose attention is mostly on the foreground (ratio > 0.5).

**Steps** only the first 30% of denoising, where layout and motion form.

- **Training-free:** no new parameters, a gain applied to the attention keys.

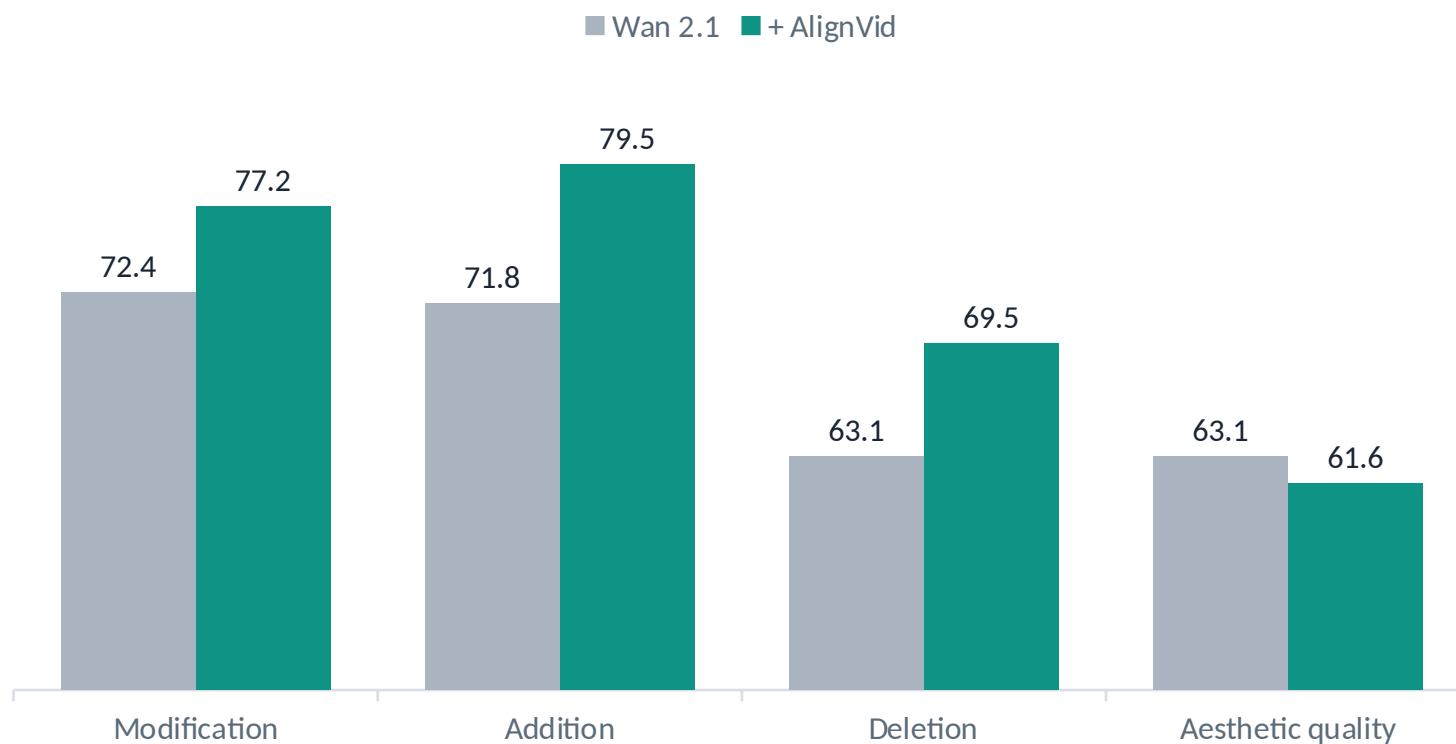
# Where scaling helps depends on the layer and the step



Each panel runs through the layers; columns are denoising steps. Blue: image tokens. Green: text. Orange: video. Below 1 means attention got sharper.

- A uniform scale wastes its effect, so we choose layers and steps.

# Up to +7.8 points of instruction following, with no training



## +7.8

Addition: 71.8 → 79.5  
the sunflower actually appears

## +6.3

Deletion: 63.1 → 69.5

## -1.5

Aesthetic score: the price

OmitI2V: 367 image-to-video edits, each checked by targeted visual questions. Higher is better.

- Much of what looks like missing capability is the model not using its inputs.

# 02

## Physical

The event happens. Does it happen the way the world works?

### PhysHPO

NeurIPS 2025 · Harold Haodong Chen et al.



Paper [arXiv:2508.10858](https://arxiv.org/abs/2508.10858) Talk video [slideslive.com/39050359](https://slideslive.com/39050359)

# One “this video is better” label can’t say what went wrong



## Wrong end state

the liquid looks wrong at the end

## Impossible motion

the path between frames breaks physics

## Missing event

the apple is never actually peeled

*Standard preference learning collapses all three into one bit per video.*

- **Baseline above, PhysHPO below in each pair. Different errors need different feedback.**

## The starting point: DPO learns from a better and a worse video

$$\Delta_{\theta} = \log \frac{p_{\theta}(y_w | x)}{p_{\text{ref}}(y_w | x)} - \log \frac{p_{\theta}(y_l | x)}{p_{\text{ref}}(y_l | x)}$$
$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \log \sigma(\beta \Delta_{\theta})$$

**y\_w, y\_l** the preferred and the rejected video for prompt x

**Δ** how much more the model prefers y\_w than a frozen reference does

**The loss** pushes Δ up; β sets how hard

- One comparison, one bit of feedback. The question is what goes into the pair.

# PhysHPO turns every real clip into four reasons

1

## Instance

### Whole clip

real vs. generated clip, or one from a partial caption

2

## State

### First & last frames

real vs. a copy with endpoints replaced

3

## Motion

### Optical flow

how things move, made explicit

4

## Semantic

### Caption

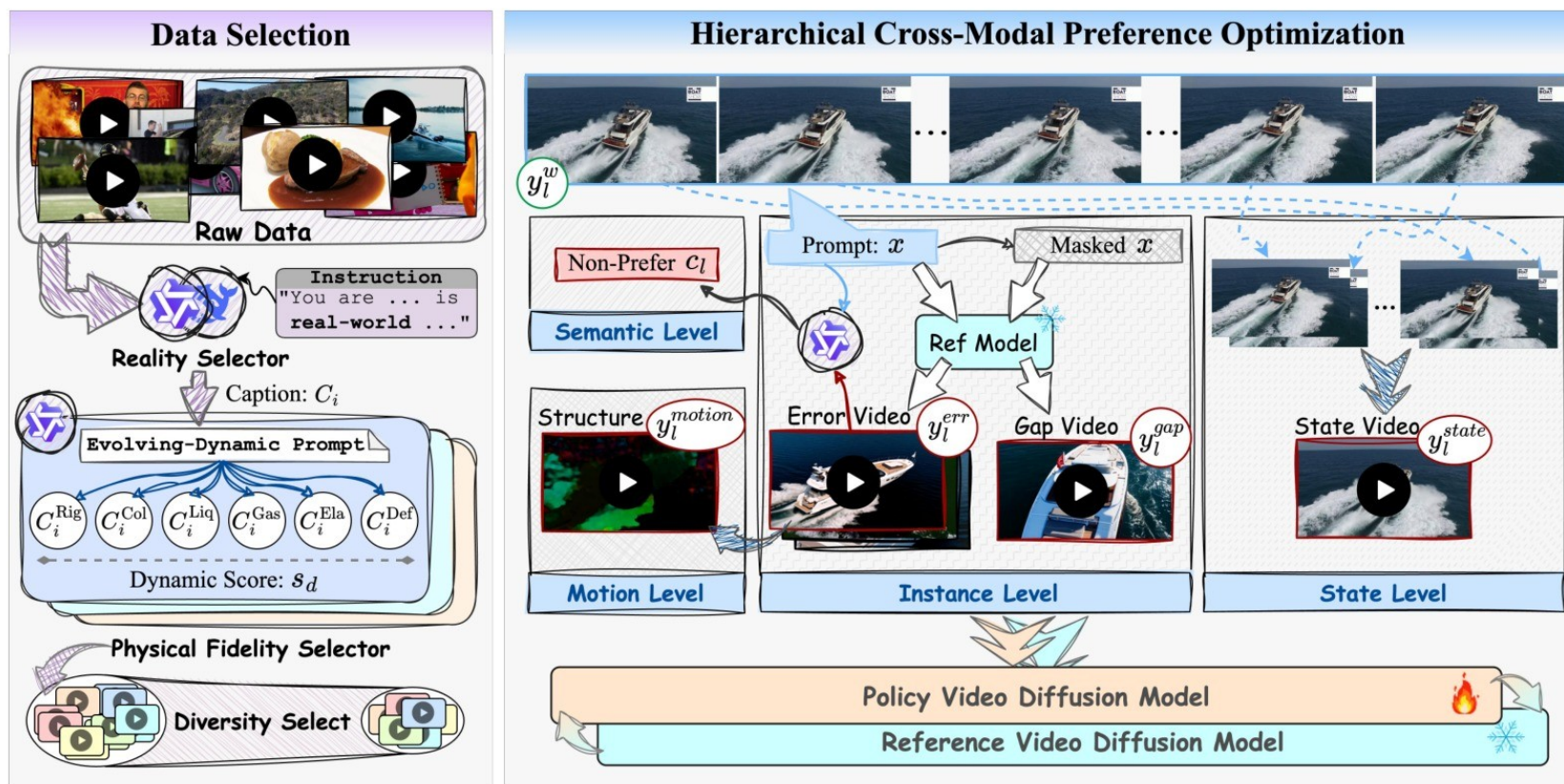
true caption vs. one describing a different event

**Why all four?** A glass tipping over can have a perfect first and last frame while it bends through the table in between. Endpoints, path and meaning each catch what the others miss.

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

- Same generator, same architecture. Each example now carries its reason.

# Select real clips, build four comparisons, train one generator



- Left: data selection. Right: the four levels. Bottom: one policy model against a frozen reference.

# Less, but more diverse, data wins

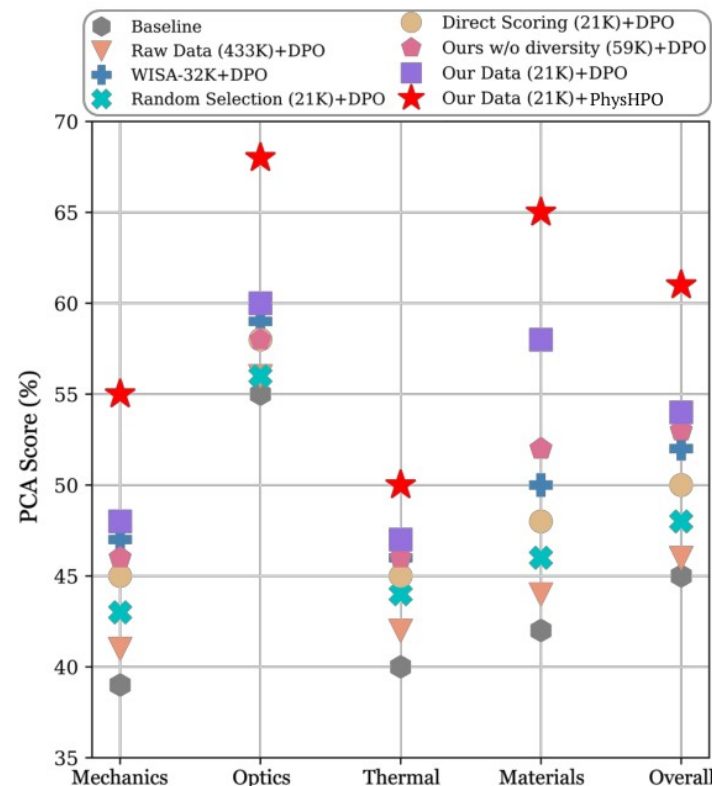
## 433K

clips in OpenVidHD

real footage → clear physical events → de-duplicated

## 21K

kept for training: about 5%



### Purple vs. pink

21K diverse clips beat 59K without the diversity step (same DPO).

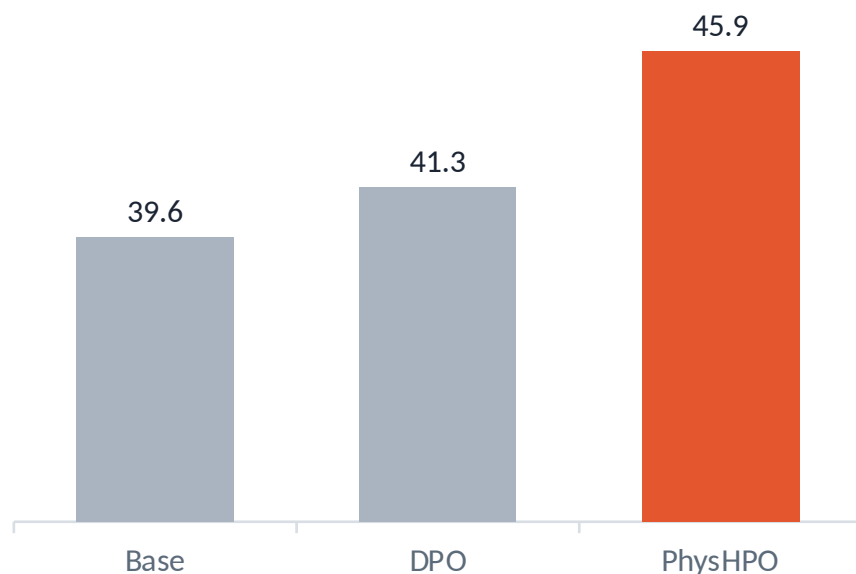
### Red stars

PhysHPO on the same 21K clips: higher in every category.

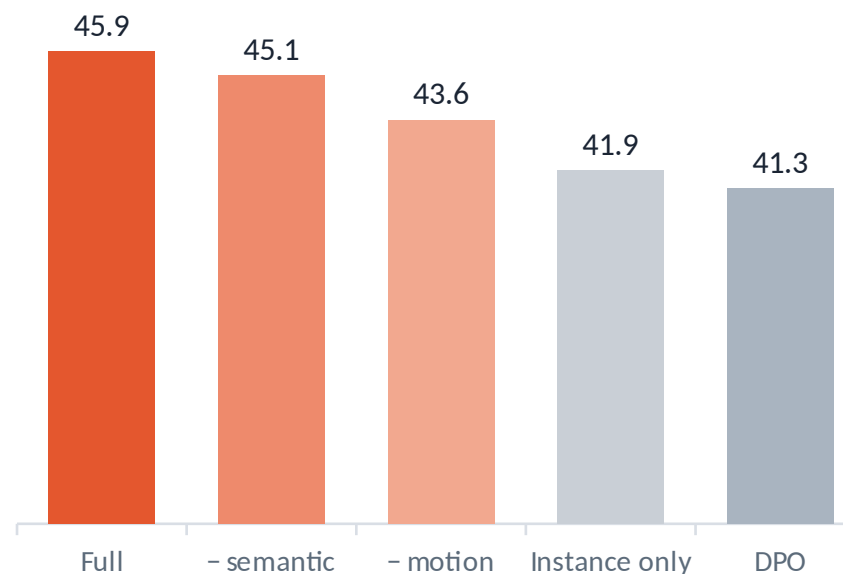
- Choosing data well helps. The richer training signal helps on top.

# +4.6 VideoPhy over DPO on the same data, and every level counts

VideoPhy (CogVideoX-5B, same data)



Ablation: remove levels in turn

**0.54 → 0.61**

PhyGenBench, DPO → PhysHPO

**82.4 → 82.8**

VBench: general quality holds

- If your data knows why a sample is bad, don't compress that into one bit.

## Next: stable online RL on a production video model

*Prompt: ...A train constructed entirely of edible delights begins its journey... From its candy-cane chimneys, it rhythmically releases vast, billowing clouds of pink marshmallow steam...*



*Prompt: ...Within the cup, a silky, creamy, warm milk foam already forms, ...Then, a delicate, steady stream of dark brown espresso, with hints of glossy caramel, is carefully poured from above...*



Training iters: 100

sampling steps: 20

- Keep exploration near the pretrained manifold, so rollouts and rewards stay trustworthy.

# 03

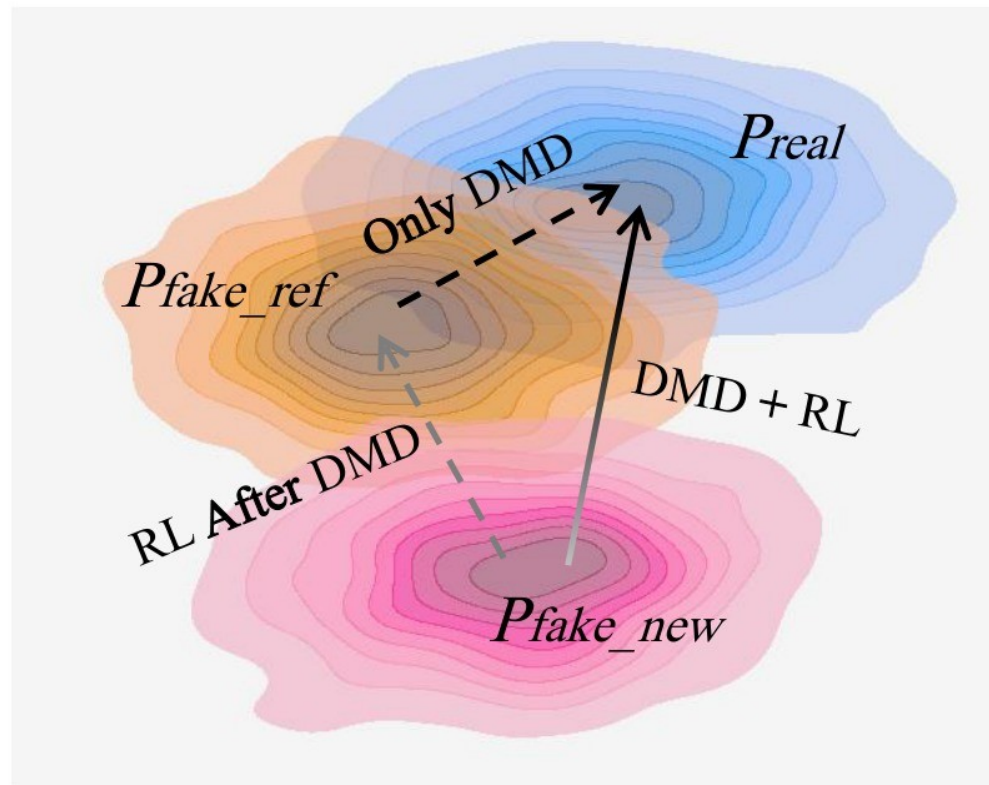
## Fast

A robot compares many futures per decision.  
Can we predict in one step?

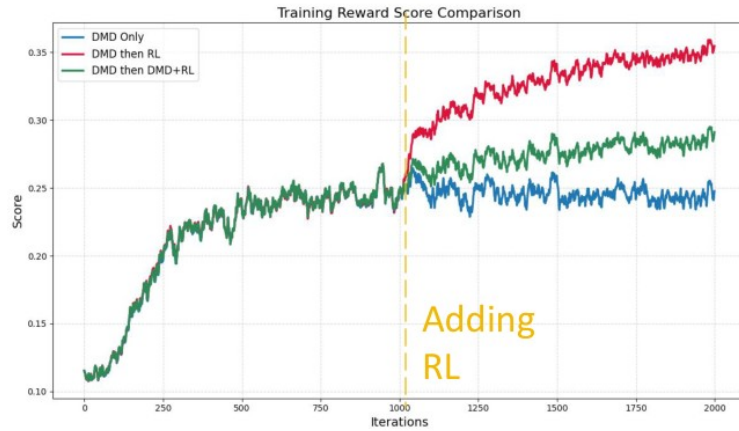
### DMDR

ECCV 2026 · Dengyang Jiang et al.

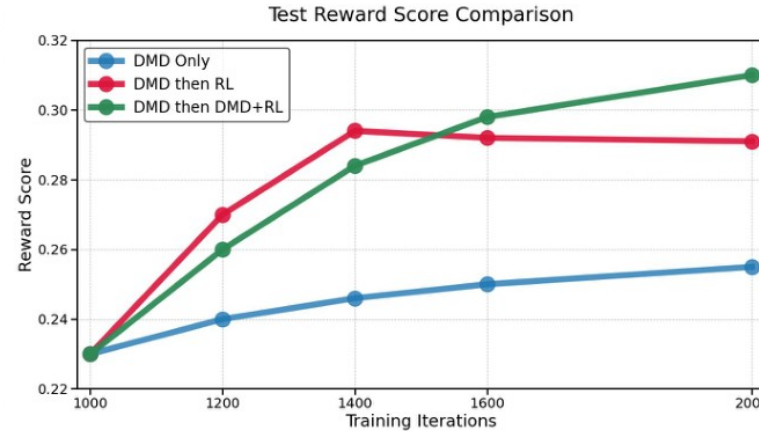
Code [github.com/vvvvjdy/dmdr](https://github.com/vvvvjdy/dmdr) Slides [hyang.org/ppt/dmdr2.pdf](https://hyang.org/ppt/dmdr2.pdf)



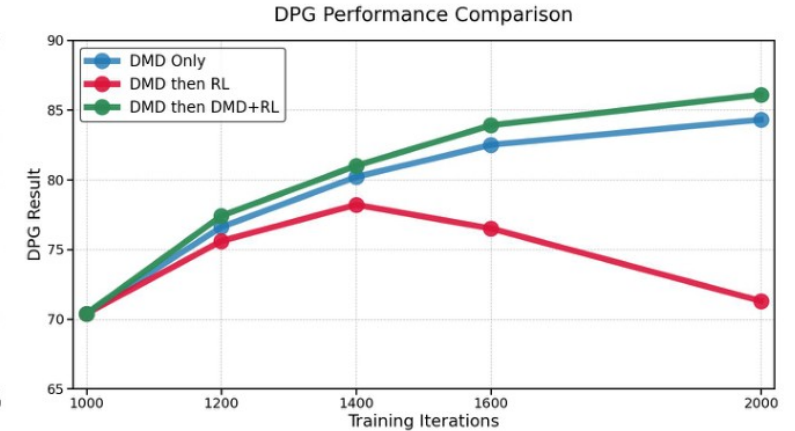
# Optimize a fast model for reward alone, and it learns to cheat



(a) Training Reward Score Comparison



(b) Test-set Reward Score Comparison



(c) DPG Score Comparison

↑ Training reward

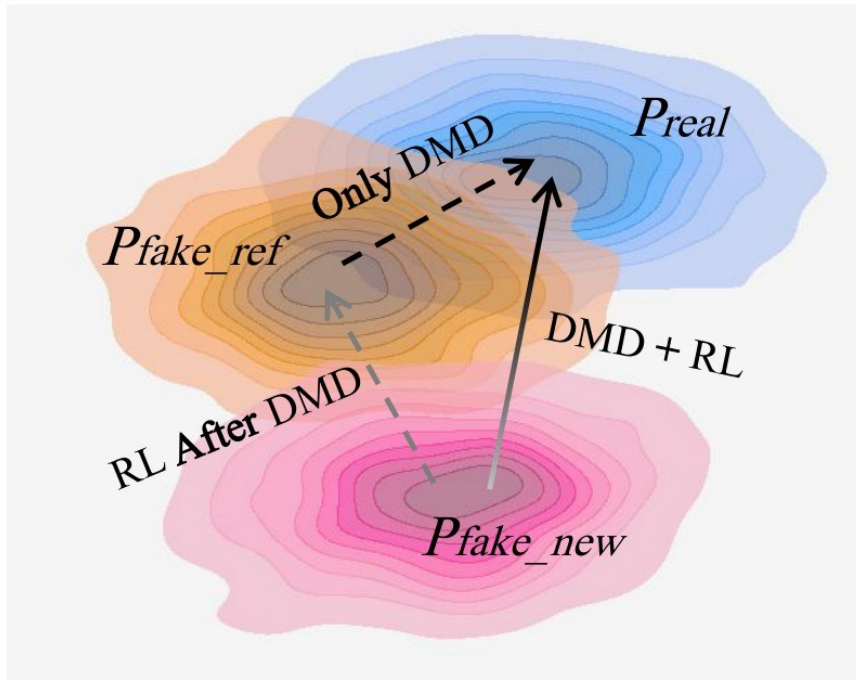
↑ Test reward

↓ Prompt following (DPG)

Red: RL after distillation. Green: DMDR keeps distillation on.

- **Reward hacking: better at pleasing the scorer, worse at the task.**

# DMDR: never stop distilling while you learn from rewards



**Distillation (DMD)** pulls the one-step student toward the teacher's distribution.

**Reward (RL)** pulls it toward what people prefer.

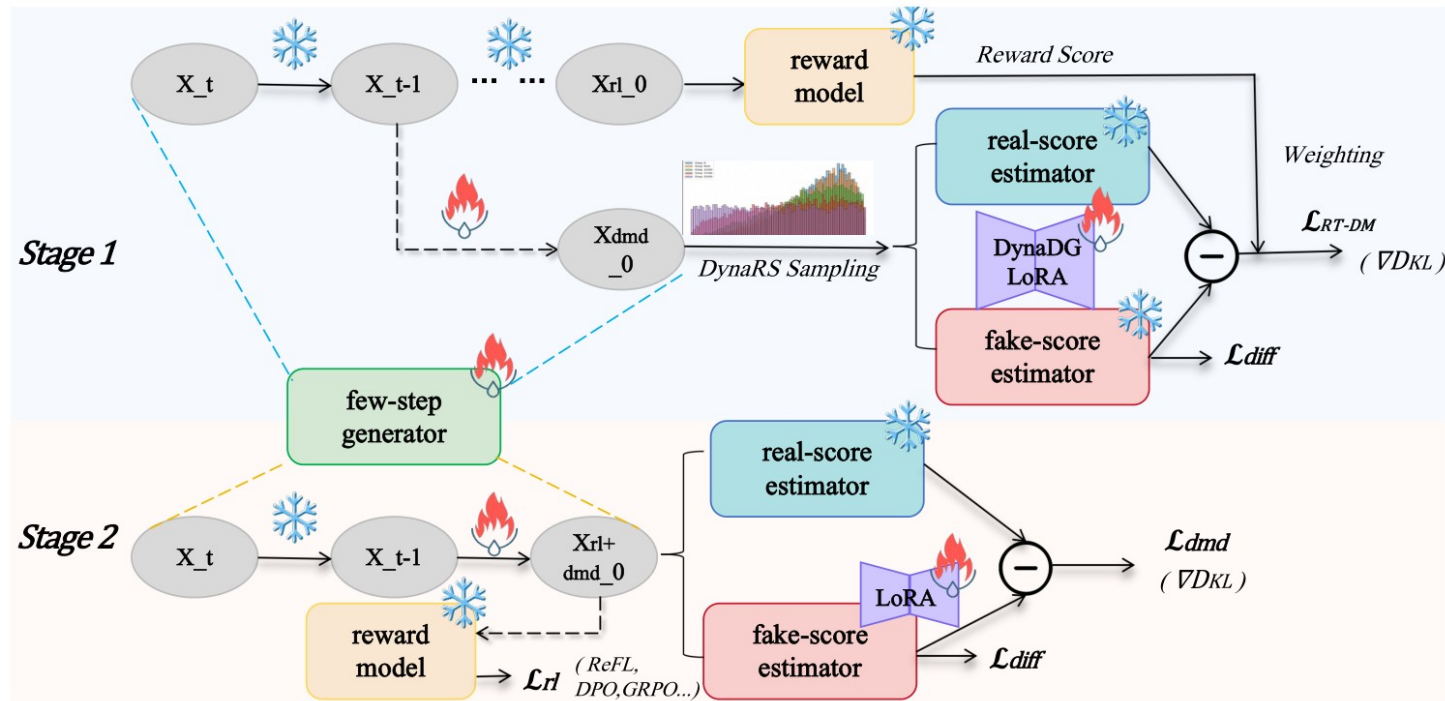
**Together** the teacher regularizes RL, so the student can't drift.

FINAL OBJECTIVE

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DMD}} + \lambda_{\text{RL}} \mathcal{L}_{\text{RL}}$$

- Keep the teacher in the loop when you optimize for reward.

# Make early training easy, then switch on direct rewards



## DynaDG

adapt the teacher-side estimator toward the student early, then anneal it away.

## DynaRS

sample high noise first, when student and teacher barely overlap.

## Stage 1

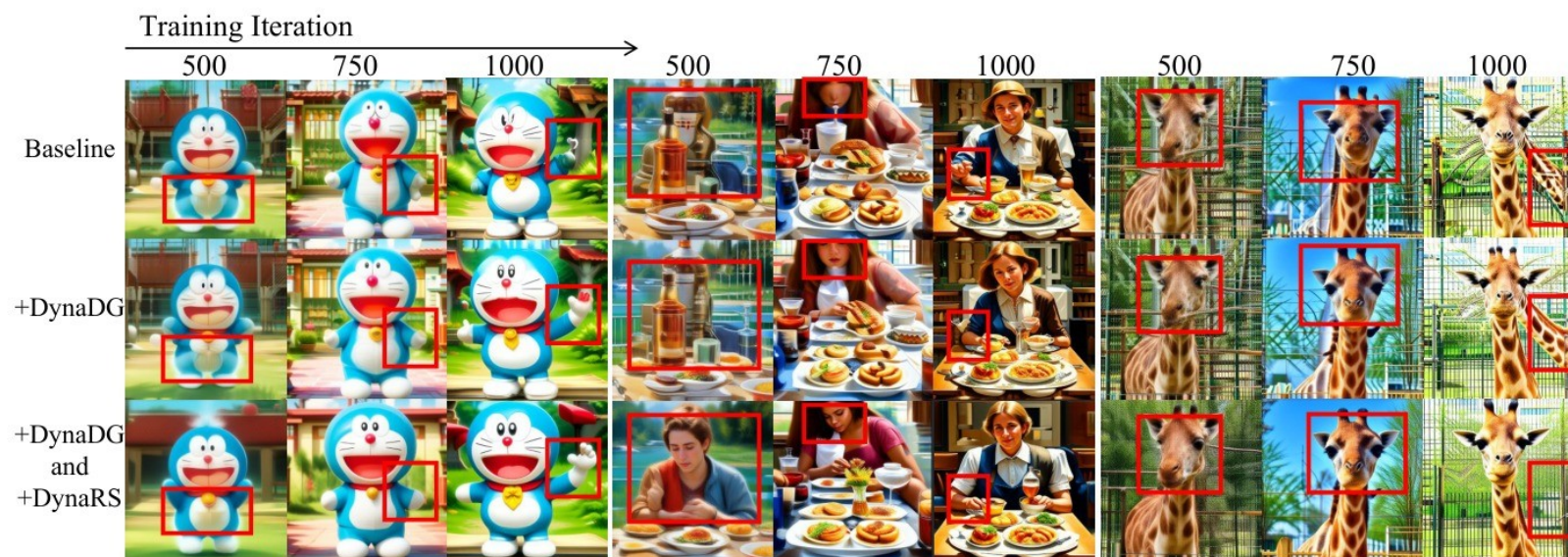
reward only reweights the distillation update.

## Stage 2

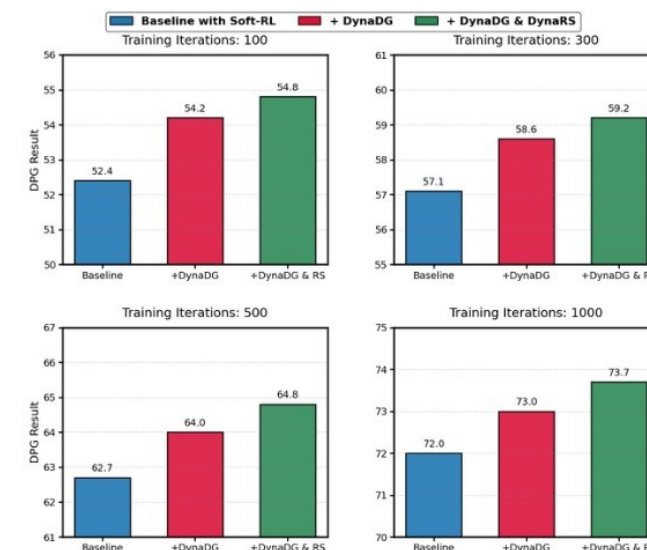
add direct reward gradients, with DMD still on.

- Snowflakes are frozen, flames are trained. Inference stays one step.

# With DynaDG and DynaRS, objects take shape earlier



(a) Visual Comparison



(b) DPG Score Comparison

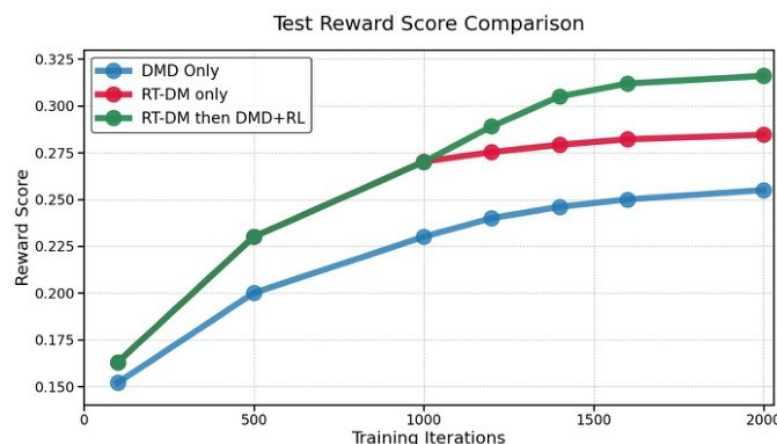
Columns: training checkpoints. Rows: baseline, + DynaDG, + DynaDG + DynaRS. Right: prompt following (DPG) at the same checkpoints.

- A better student, sooner: the second stage needs one.

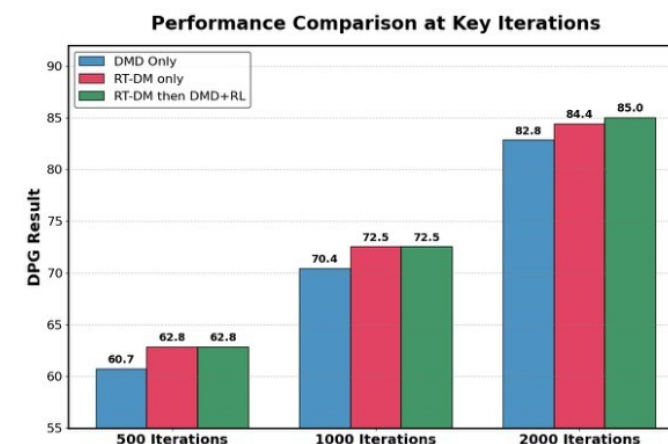
# Switching on direct rewards keeps improving after progress stalls



(a) Training Reward Score Comparison



(b) Test-set Reward Score Comparison



(c) DPG Score Comparison

Blue DMD only

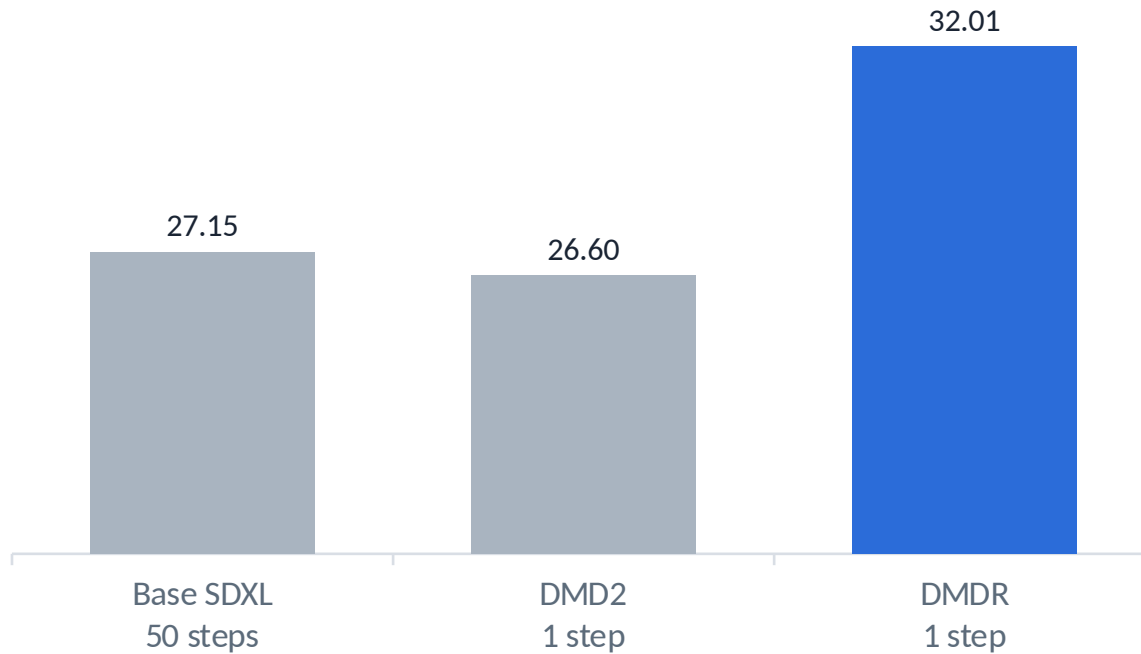
Red reward-weighted DMD throughout

Green adds direct reward in stage 2

- Green climbs on training reward, test reward and prompt following together.

# One step beats fifty

HP score (higher is better)



## 32.01

HP score at 1 step, vs. 26.60 for DMD2

## 35.62

CLIP score at 1 step, vs. 34.76 for the 50-step base

SDXL backbone. Steps count generator evaluations, not total system latency.

- With the right feedback, speed doesn't have to cost quality.

## DMDR helped distill Z-Image-Turbo, an open 8-step model



- Samples from the released model, distilled with Decoupled-DMD and DMDR.

# For video, speed also means 4-bit attention

FP16



INT4 (ours)



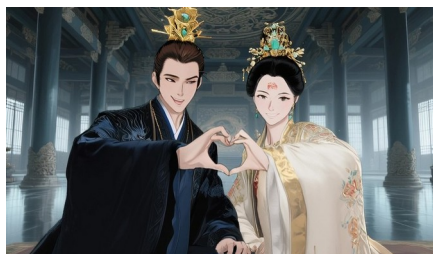
# 1.68×

end-to-end speedup on Wan 2.2



# 2.2×

faster attention kernel



# 81%

of attention calls run in INT4 (Wan 2.2 text-to-video)

- Handle 3D-RoPE outliers and use the full 4-bit range: quality stays nearly identical.

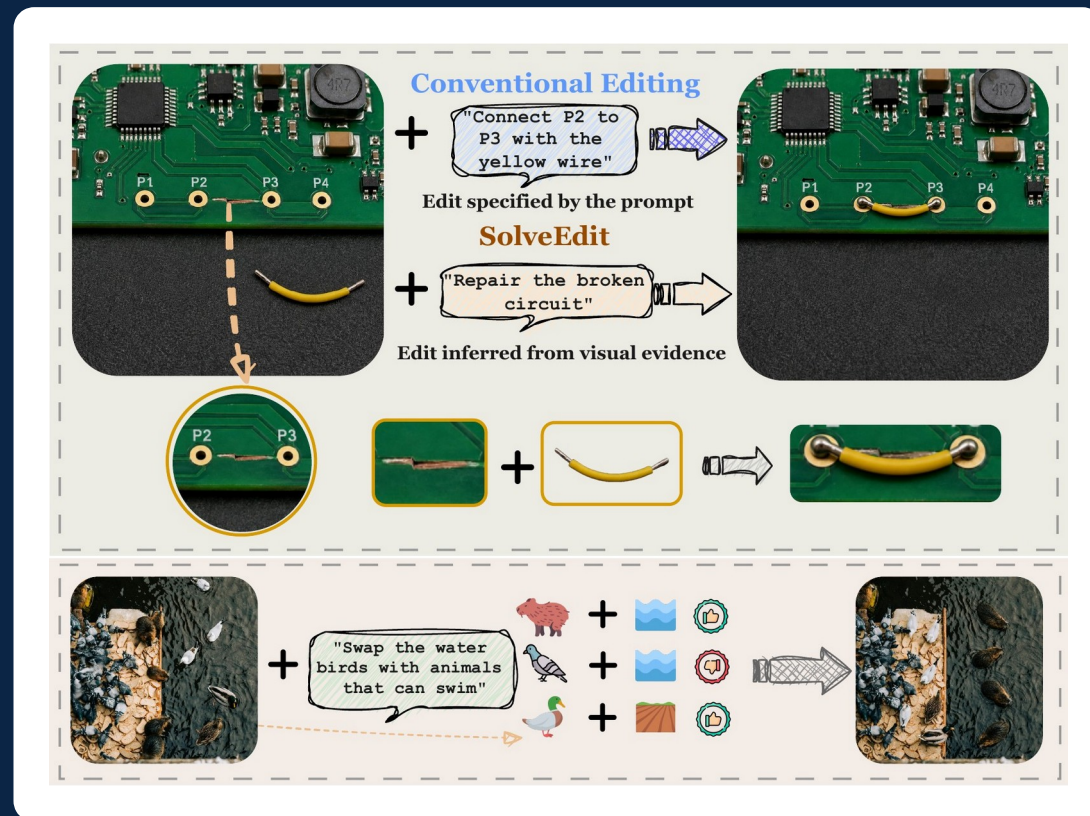
# 04

## Goal-directed

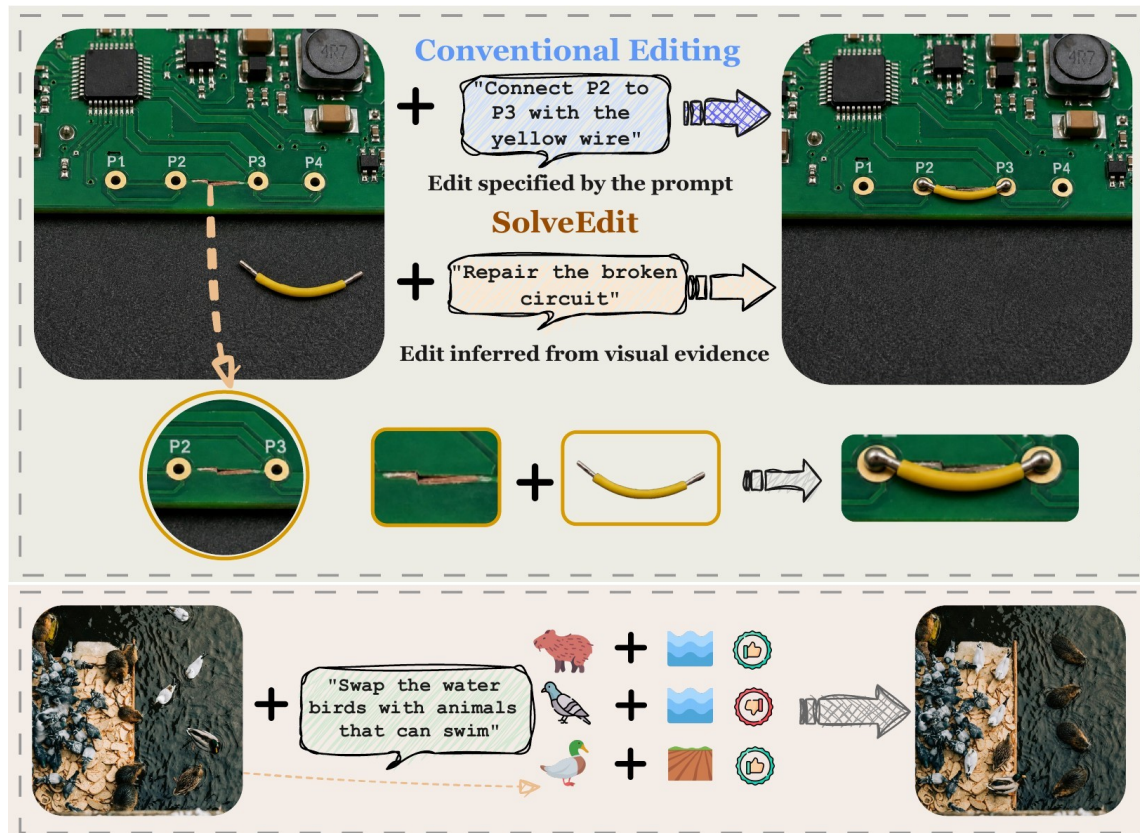
Give the model a goal, not an edit. Can it work out what has to change?

### SolveEdit

arXiv 2026 · Wen-Jie Shu, Yexin Liu et al.



# An editor is told the change. A solver must find it



Editing "Connect P2 to P3 with the yellow wire"

Solving "Repair the broken circuit"

Find the break, pick the right part, and change nothing else.

- Same output. Very different amount of thinking.

# 2,728 problems, grouped by where the answer comes from

## The prompt says it

Instruction-specified



Remove the fallen debris.



## The scene says it

State-dependent



Which segment repairs the road over the river?

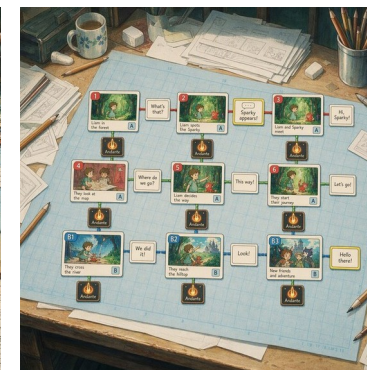


## A rule in the image says it

Rule-dependent



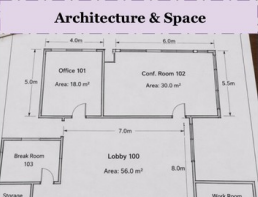
Where does each storyboard card belong?



Each pair: input, then GPT-Image-2's output. 10 domains, 54 subdomains.

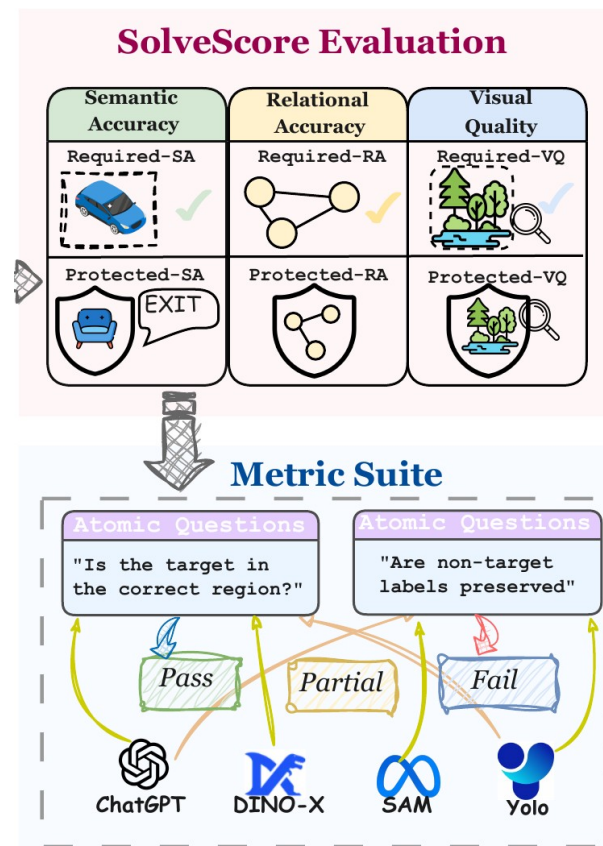
- Many outputs can be correct, so there is no single reference image.

# Ten domains, from molecules to card games

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Science &amp; Technology</b></p>  <p><b>Prompt:</b><br/>Complete the molecular model using the parts from the tray so that it matches the provided diagram.</p> <p><b>POS</b> Is one white sphere attached to the far-right black sphere of the five-membered ring?</p> <p><b>ANS</b> Yes. A white sphere is attached to the right-most black sphere on the 5-membered ring.</p> <p><b>Partial:</b> Wrong-colored (non-white) atom attached, or white sphere loosely placed without a connecting stick.</p> <p><b>Fail:</b> No sphere is added to the far-right black sphere.</p> <p><b>NEG</b> Is the tablet display free of blur, warping, and smearing?</p> <p><b>ANS</b> Yes. The tablet screen display is sharp, clean, and free of generation artifacts.</p> <p><b>Partial:</b> Slight blurring or minor smearing is visible on the tablet screen.</p> <p><b>Fail:</b> The tablet screen exhibits severe blurring, warping, or heavy smearing.</p> | <p><b>People &amp; Daily Life</b></p>  <p><b>Prompt:</b><br/>Organize the office workspace by relocating the binders to a different compartment on the shelving unit.</p> <p><b>POS</b> Is the original location on the lower shelf free of ghosting or erasing artifacts?</p> <p><b>ANS</b> Yes. The row of binders is clearly present in the output image.</p> <p><b>Partial:</b> Most of the binders are present, but a few are missing.</p> <p><b>Fail:</b> The row of binders is completely absent from the output image.</p> <p><b>NEG</b> Is the man standing in the background preserved intact?</p> <p><b>ANS</b> Yes. The man standing in the background is preserved.</p> <p><b>Partial:</b> The man is mostly preserved but has minor unnatural alterations.</p> <p><b>Fail:</b> The man is missing, warped, blurred incorrectly, or inappropriately modified.</p> | <p><b>Nature &amp; Wildlife</b></p>  <p><b>Prompt:</b><br/>Reorganize the woodland gathering by moving small and young animals to a separate area away from large mammals.</p> <p><b>POS</b> Is there a clear spatial gap separating the new cluster from the moose, bear, and robot?</p> <p><b>NEG</b> Do the moose, bear, and robot remain present and visually intact?</p> <p><b>Object &amp; Artifact</b></p>  <p><b>Prompt:</b><br/>Show the watch twenty-six hours later.</p> <p><b>POS</b> Does the date window display the number 24?</p> <p><b>NEG</b> Do the watch-face geometry, date-window location, and hand pivot remain aligned with their original positions?</p> <p><b>Architecture &amp; Space</b></p>  <p><b>Prompt:</b><br/>Correct the labels on the architectural floor plan to be consistent with the visible dimensions and rules.</p> <p><b>POS</b> Does the area label for Conf. Room 102 read exactly Area: 33.0 m<sup>2</sup>?</p> <p><b>NEG</b> Are the ruler, pencil, table, and architectural lines preserved?</p> | <p><b>Art &amp; Design</b></p>  <p><b>Prompt:</b><br/>Move the papers and design items into the correct WARM, COOL, NEUTRAL, and ACCENT zones according to the legend and dominant color family.</p> <p><b>POS</b> Are all seven accent items fully enclosed within ACCENT?</p> <p><b>NEG</b> Are the grid, four zone labels, and RULES card intact and legible?</p> <p><b>Game &amp; Puzzle</b></p>  <p><b>Prompt:</b><br/>Play a valid card from hand into the corresponding item slot, satisfying its attribute requirements.</p> <p><b>POS</b> Does it match one of the six original cards: A♥, K♠, Q♣, 10♥, 9♠, or 3♥?</p> <p><b>ANS</b> Yes. The deployed card matches the suit and rank of one of the original cards from the player's hand.</p> <p><b>Partial:</b> Suit and rank mostly match a source card, though symbols are slightly unclear.</p> <p><b>Fail:</b> The card doesn't match any original hand card, or its identity is unreadable.</p> <p><b>NEG</b> Are the boss and its health bar unchanged?</p> <p><b>ANS</b> Yes. Boss character and health bar preserved identically.</p> <p><b>Partial:</b> Minor modifications to boss or health bar, but values largely intact.</p> <p><b>Fail:</b> Boss/health bar missing or content/value clearly altered.</p> | <p><b>General</b></p>  <p><b>Prompt:</b><br/>Completely empty every container holding a secondary-colored liquid.</p> <p><b>POS</b> Is cup 2 completely empty, with no orange liquid remaining?</p> <p><b>ANS</b> Yes. The 2nd cup from the left is completely empty, showing clear plastic with no orange liquid inside.</p> <p><b>Partial:</b> The 2nd cup has much less orange liquid, but isn't empty or retains a faint tint.</p> <p><b>Fail:</b> The 2nd cup still contains a full or nearly full amount of orange liquid.</p> <p><b>NEG</b> Are cups 1, 3, and 5 and their red, yellow, and blue liquids unchanged?</p> <p><b>ANS</b> Yes. Exactly six cups present, background and surface textures untouched.</p> <p><b>Partial:</b> Six cups are present, but there are minor alterations or smudges to the background or surface textures.</p> <p><b>Fail:</b> Number of cups changed, or visible alterations to background or surface.</p> |
| <p><b>Sports &amp; Motion</b></p>  <p><b>Prompt:</b><br/>Draw all expected forward trajectory routes for the active airborne entities to complete the visible play.</p> <p><b>POS</b> Does a trajectory start at defender 25 and point downward toward the court?</p> <p><b>NEG</b> Are the crowd and court preserved without semantic changes?</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <p><b>Geography &amp; Environment</b></p>  <p><b>Prompt:</b><br/>Mark all valid pedestrian routes following the constructed trail through the landscape.</p> <p><b>POS</b> Does the route extend to the end of the trail near the valley vegetation?</p> <p><b>NEG</b> Is the natural landscape preserved without blur or distortion?</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |

- Each case: a goal, required checks (POS) and protected checks (NEG), with pass, partial and fail criteria.

# Score what must change, and what must stay



## Required

"Is the target in the correct region?"

## Protected

"Are the non-target labels preserved?"

$$\text{SolveScore} = G_{\text{quality}} \cdot \max(0, R - \lambda D)$$

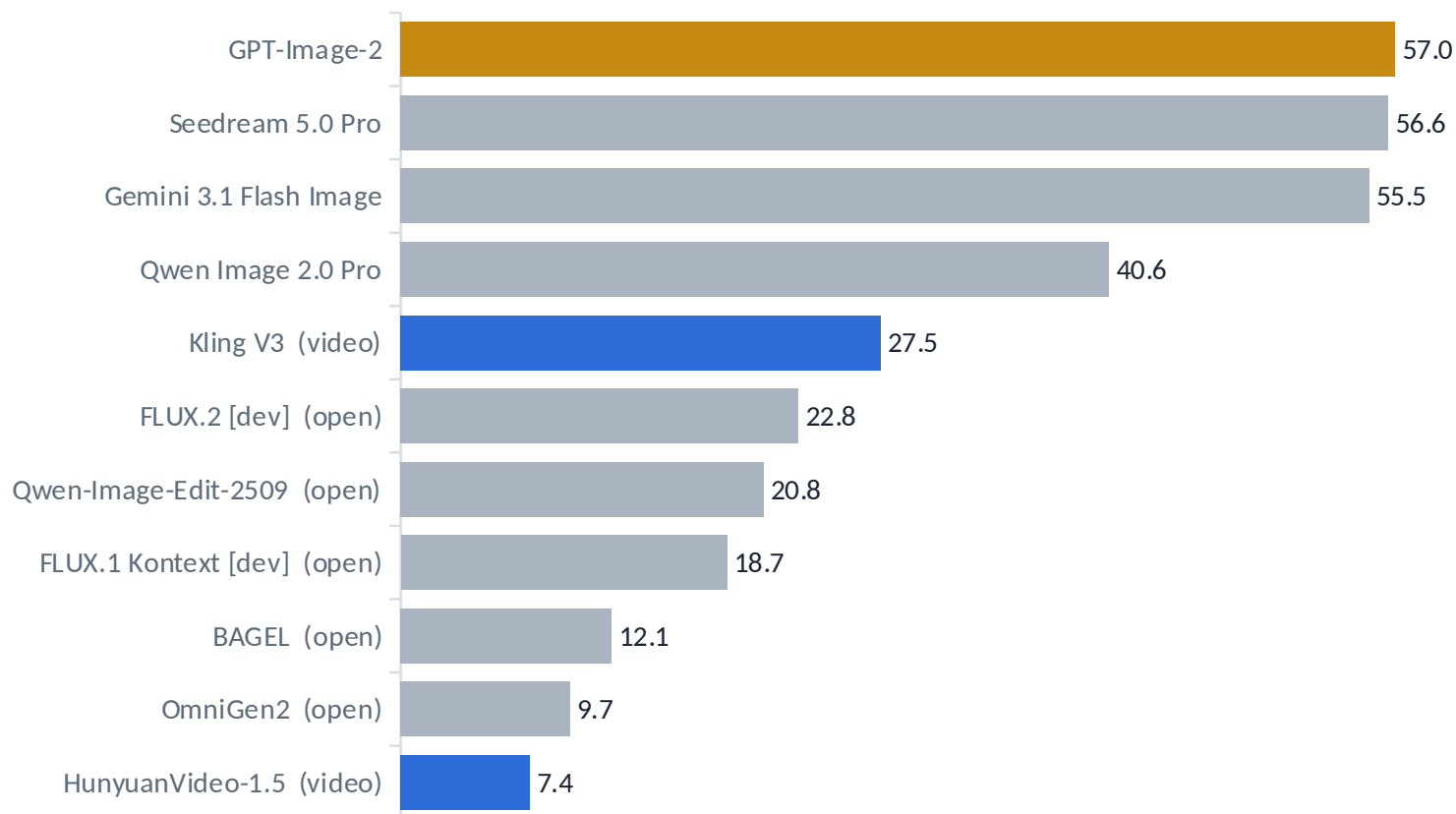
$$R = \sum_{q \in C_{\text{req}}} w_q a_q \quad D = \sum_{q \in C_{\text{pro}}} u_q (1 - a_q)$$

**R** = required change achieved    **D** = protected content damaged    **λ = 0.5**

A VLM judges each yes/no check; SAM and YOLO add evidence where VLMs are unreliable.

- A contract instead of a reference: completion minus collateral damage.

# The best model solves 57%. Video models, far less



## 57.0

best SolveScore, GPT-Image-2

## 22.8

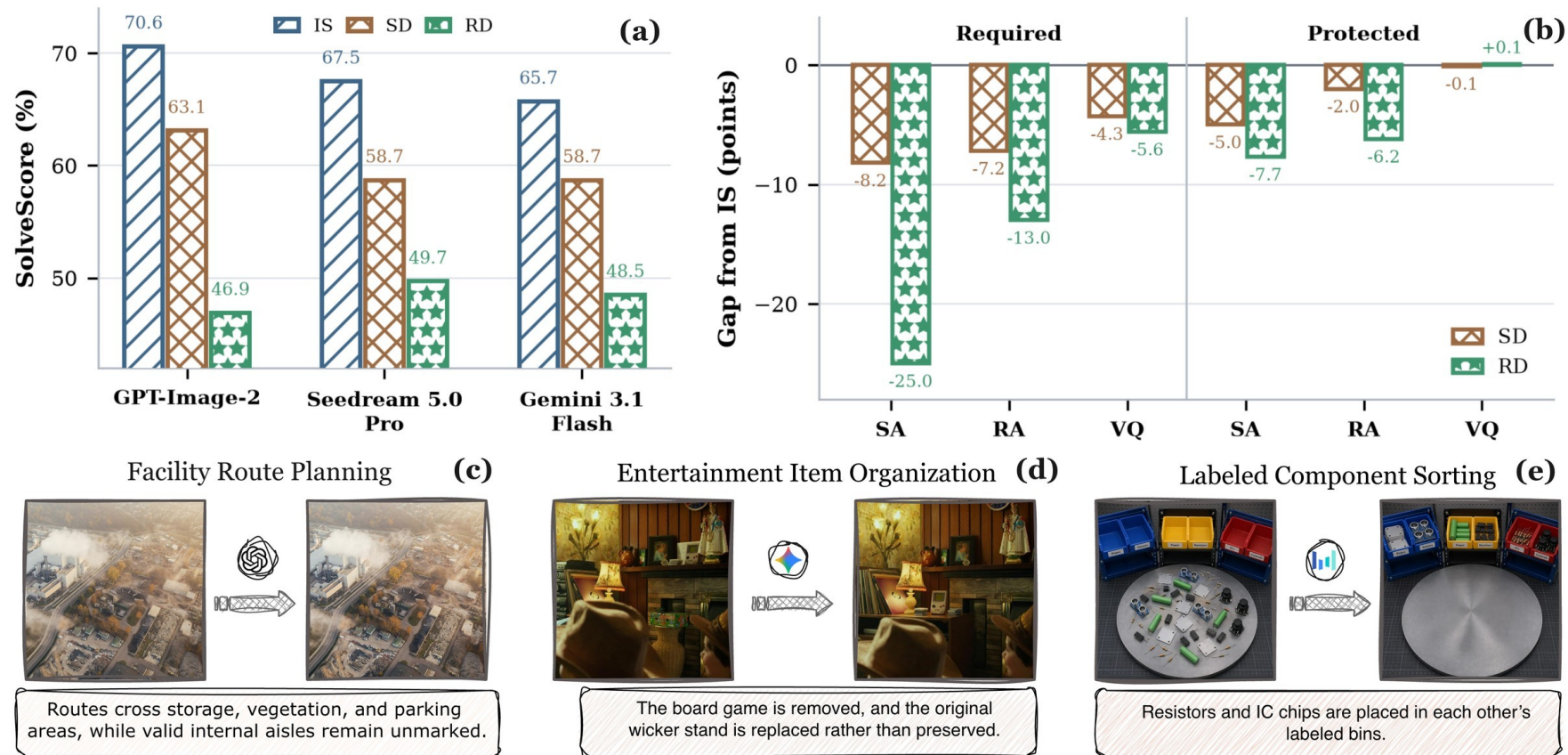
best open-weight editor

## 7.4

HunyuanVideo-1.5, scored on its final frame

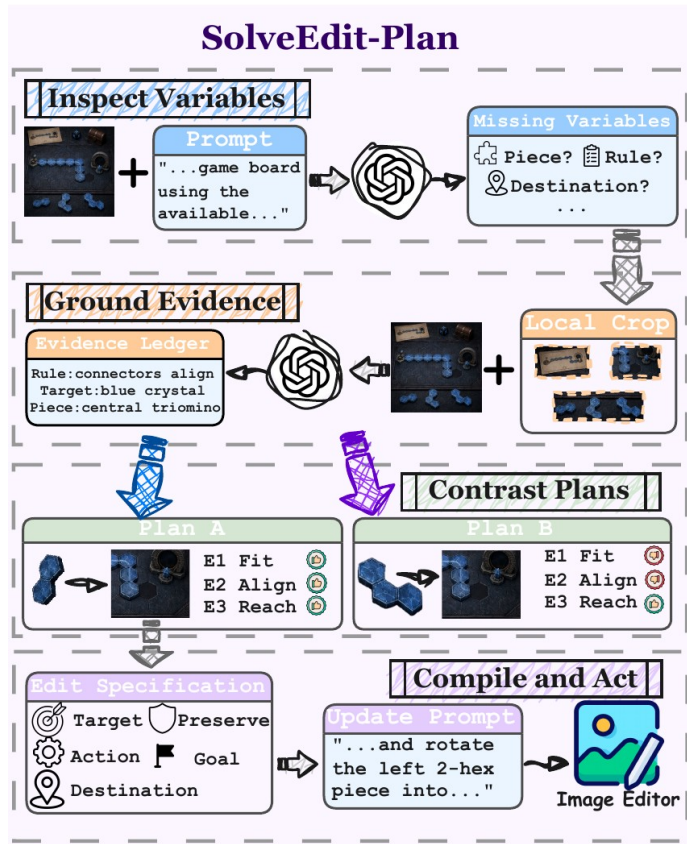
● For video models to become world models, this is the gap to close.

# The outputs look right. The problem was solved wrong



- Rule-dependent cases drop 17-24 points, in meaning and relations, not visual quality.

# The fix: decide the transition before generating



1

## Inspect

Find what the request leaves open (target, rule, destination), and crop the image for evidence.

2

## Resolve

Compare candidate plans against the evidence, pick one, and write an explicit instruction that also says what to preserve.

3

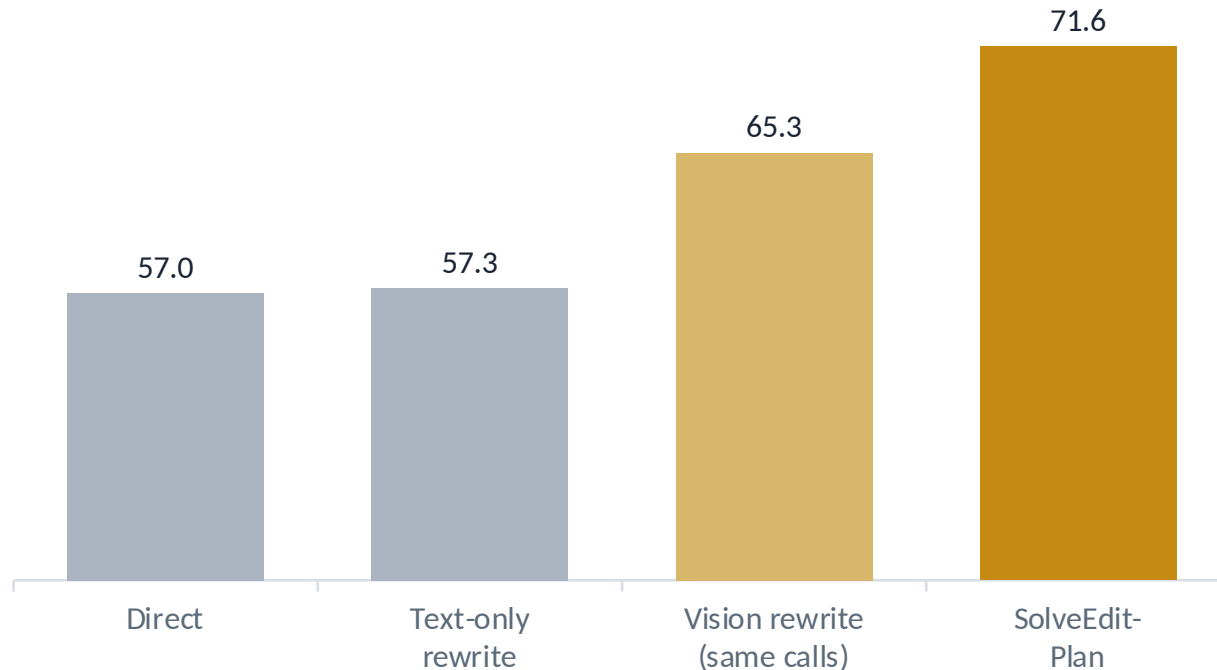
## Generate

One call to the unchanged editor. No training.

- The planner never sees the contract, the reference, or the case type.

# Planning first: 57.0 → 71.6, with half the collateral damage

GPT-Image-2, one final generation each



**16.3 → 8.6**

collateral damage vs. the matched vision rewrite

**46.9 → 64.7**

hardest cases: rule-dependent

**7.4 → 14.0**

same planner on HunyuanVideo-1.5

- For a robot, collateral damage is knocking over the bottle while reaching for the cup.

## Each paper fixes one link in the loop



- Goal → plan → imagine → compare → act: the loop a world model has to close.

## Three open problems we are working on



### Action-conditioned prediction

Robot actions, not just text, as the control signal.



### Closed-loop evaluation

Fix the planner, swap the world model, and measure task success, not video quality.



### Knowing what it doesn't know

Represent uncertainty instead of confidently rendering one guess.

- The real test of a world model: does a fixed planner succeed more often with it?

## FOUR TAKEAWAYS

01

CONTROLLABLE

Check the model is using its inputs before you train it.

02

PHYSICAL

Give feedback that says why, not just which.

03

FAST

Keep the teacher in the loop when you optimize for reward.

04

GOAL-DIRECTED

Decide what should change before you render it.

# Thank you

Questions welcome.

## WORK LED BY

● **Yexin Liu**  
AlignVid, SolveEdit

● **Mingzhe Zheng**  
SAGE-GRPO

● **Yaofu Liu**  
RotateAttention

● **Harold Haodong Chen**  
PhysHPO

● **Dengyang Jiang**  
DMDR

● **Wen-Jie Shu**  
SolveEdit

...with collaborators including Ser-Nam Lim, Qifeng Chen, Xiaogang Xu and Steven Hoi.



**hyang.org**

papers • code • slides • demos

DNA Lab • HKUST

## Why scaling logits always lowers entropy

$$p_j(\alpha) = \frac{e^{\alpha z_j}}{\sum_k e^{\alpha z_k}}$$

$$\frac{dH}{d\alpha} = -\alpha \operatorname{Var}_{p(\alpha)}(z) \leq 0$$

$$\frac{d}{d\alpha}(\log Z - \alpha \mathbb{E}_p[z]) = \mathbb{E}_p[z] - \mathbb{E}_p[z] - \alpha \operatorname{Var}_p(z)$$

- Sharpening is guaranteed. Sharpening the right tokens is what scheduling is for.

# The DPO and DMD objectives behind PhysHPO and DMDR

**DPO** (PhysHPO builds on it)

$$\Delta_{\theta} = \log \frac{p_{\theta}(y_w | x)}{p_{\text{ref}}(y_w | x)} - \log \frac{p_{\theta}(y_l | x)}{p_{\text{ref}}(y_l | x)}$$
$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \log \sigma(\beta \Delta_{\theta})$$

**DMD gradient** (DMDR builds on it)

$$\nabla_{\theta} \mathcal{L}_{\text{DMD}} \approx -\mathbb{E} \left[ w(t) (s_T(x_t, t) - s_{\theta}(x_t, t)) \frac{\partial G_{\theta}(z)}{\partial \theta} \right]$$

- **DPO:** prefer the good video more than the reference does. **DMD:** teacher score minus student score.

## Scoring one SolveEdit case

**4 required checks**

pass, pass, partial, fail

**R = 0.625**

**2 protected checks**

pass, partial

**D = 0.25**

**SolveScore**

$\max(0, 0.625 - 0.5 \times 0.25)$

**= 0.50**

- 29,460 checks in total, about 11 per case; every case has at least one protected check.

# What each comparison controls for

-  **AlignVid** Same trained model, with and without attention scaling.
-  **PhysHPO** Same selected data and base model; only the preference signal differs.
-  **DMDR** Same number of generator steps (1). Wall-clock speed needs runtime measurements.
-  **SolveEdit-Plan** Same LLM, call budget and editor as the generic vision rewrite: structure adds 6.3 of the 14.6 points.

# Papers, demos and code

- **AlignVid** ICML 2026 [project page](#) · [code](#) · [OmitI2V](#)
- **PhysHPO** NeurIPS 2025 [arXiv:2508.10858](#) · [talk video](#)
- **SAGE-GRPO** arXiv 2026 [video gallery](#) · [code](#)
- **DMDR** ECCV 2026 [code](#) · [slides](#) · [Z-Image-Turbo](#)
- **RotateAttention** ECCV 2026 [arXiv:2607.02584](#) · [code](#)
- **SolveEdit** arXiv 2026 [project page](#) · [leaderboard](#) · [code](#)

Everything else: [hyang.org](https://hyang.org)