Qualcomm AI research

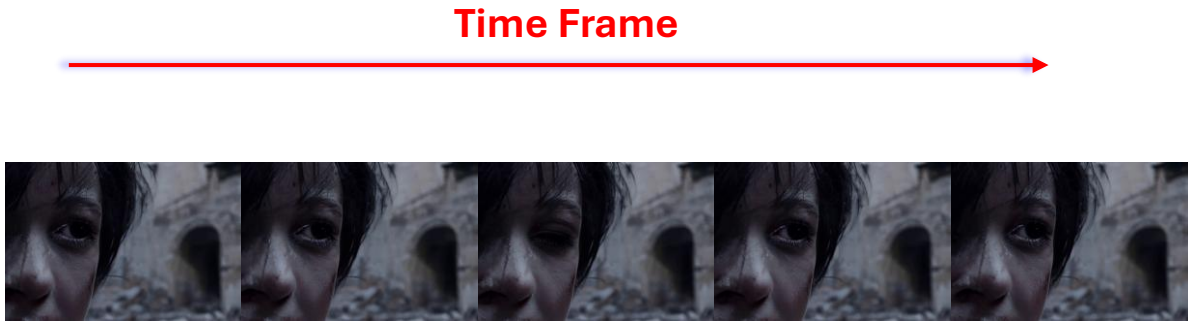
ICLR

# MoAlign: Motion-Centric Representation Alignment for Video Diffusion Models

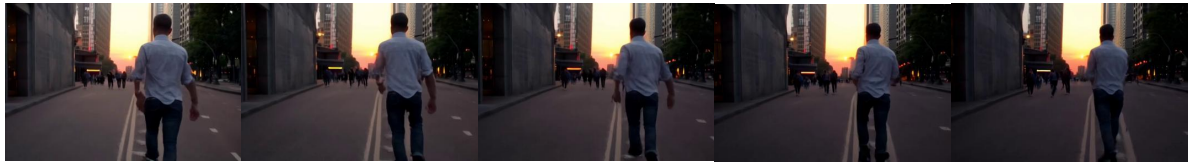
Aritra Bhowmik, Denis Korzhenkov, Cees G. M. Snoek,
Amir Habibian, Mohsen Ghafoorian

# Video Diffusion Models: Beautiful but Physically Flawed

- Generate high-quality, photorealistic frames
- Capture rich appearance & style
- But… motion often lacks **physical plausibility**



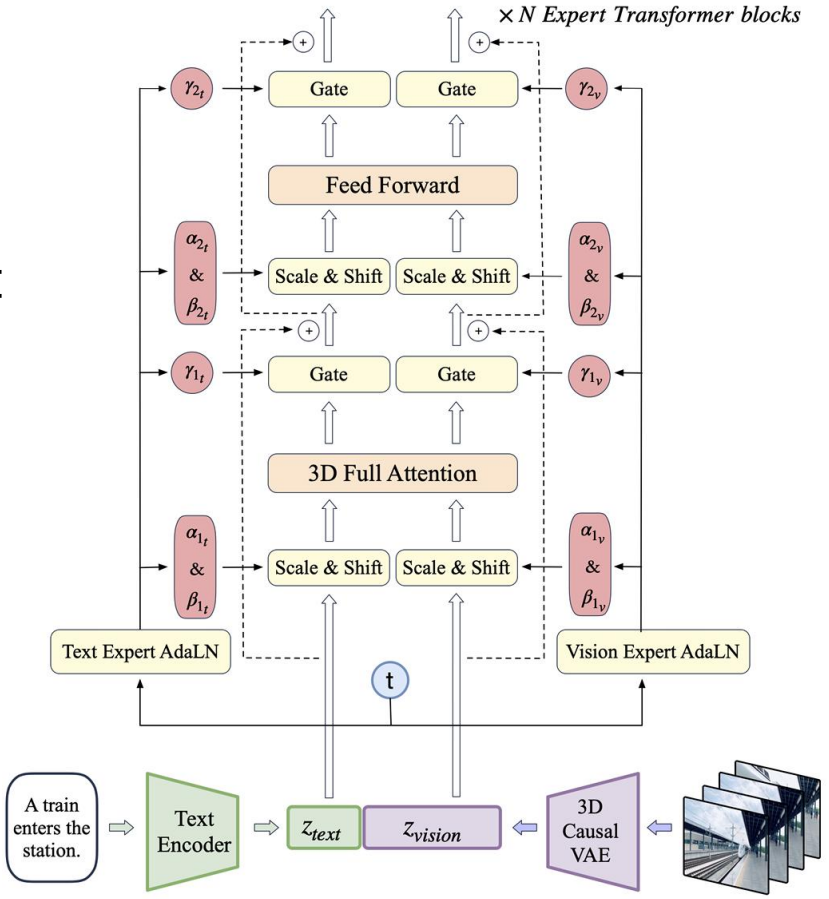

**Prompt:** Zoomed-in face of a girl staring intensely



**Prompt:** A man walking forward in a city street

# VDMs optimize for appearance, not motion

- VDMs generate videos by **denoising noisy latents step by step**
- Training focuses on **pixel/latent reconstruction**, guided by text
- Objective ensures **frame-level realism** and semantic alignment
- But…
  - No explicit term for **motion dynamics**
  - Physics laws (gravity, collisions, fluids) are **not supervised**

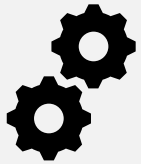



**Training Objective:** $$L_{\text{diff}} = \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\right]$$

Where $z_t$ : noisy latent video, $\epsilon$ : Gaussian noise, $c$ : text noise
Model learns to predict noise

# Three strategies have emerged to improve motion

## Simulation Based



- Physics Engines
- Simulators

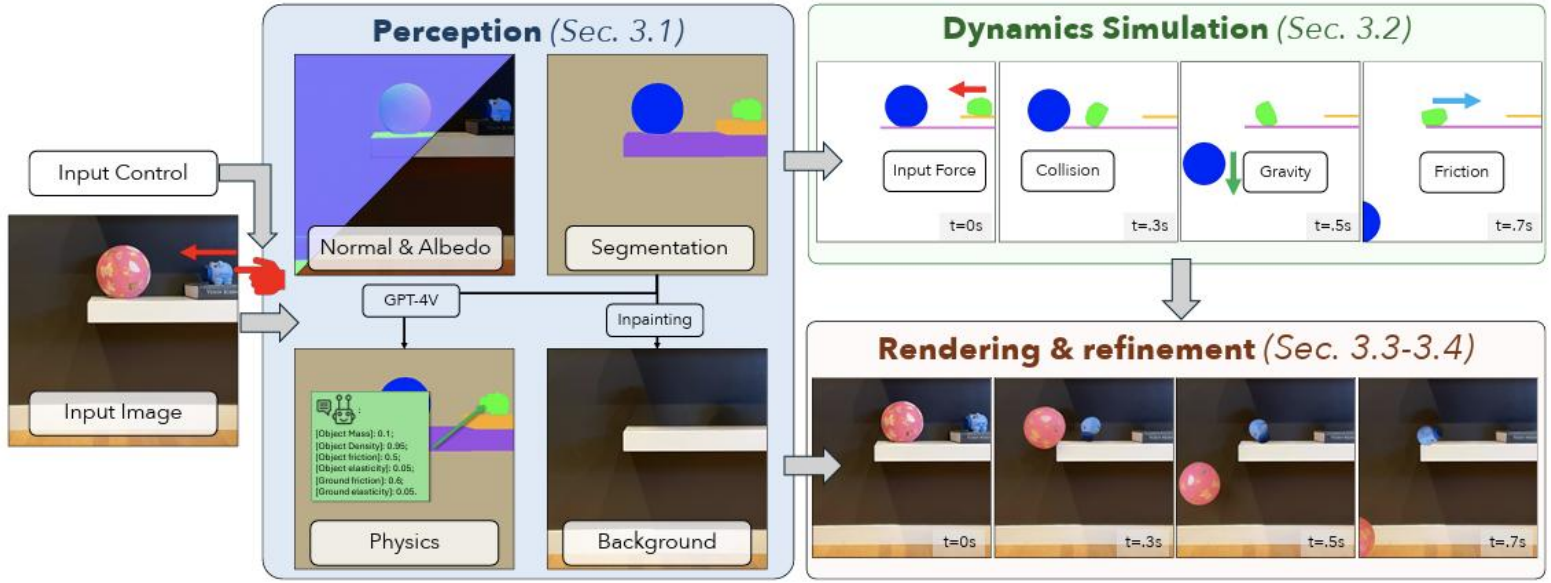
## Condition-based

- Trajectory input
- Optical flow condition

## Representation Alignment

- Align features with self-supervised video encoders

# Simulation-based methods enforce motion via physics




- Infers **object properties** (geometry, materials) from a single input image
- Runs a **rigid-body physics simulation** to generate object trajectories
- Uses **diffusion refinement** to render realistic videos from simulated motion

*Simulation ensures physical plausibility; diffusion ensures realism.*

PhysGen: Rigid-Body Physics-Grounded Image-to-Video Generation, ECCV 2024

# These methods lack realism, scalability, diversity

- **Limited scalability —** works for rigid bodies, but struggles with complex, deformable, or open-world scenes
- **Visual realism gap —** simulators generate plausible motion, but outputs look synthetic without heavy refinement
- **High cost & rigidity —** requires explicit physics setup and is not flexible for creative or diverse prompts

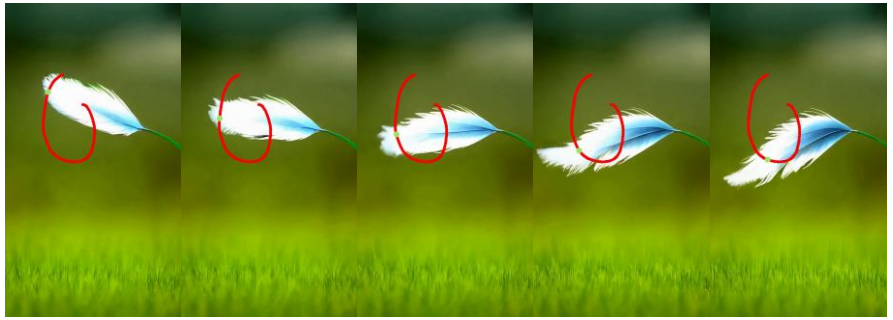
*Physics engines enforce laws of motion but limit realism, scalability, and generality.*

PhysGen: Rigid-Body Physics-Grounded Image-to-Video Generation, ECCV 2024

# Trajectory conditioning enables precise motion control

- Input: **text + trajectory path**
- Model fuses trajectory with text to guide motion
- Capable of generating **videos that follow arbitrary paths**

First frame shows the input trajectory for the generated video

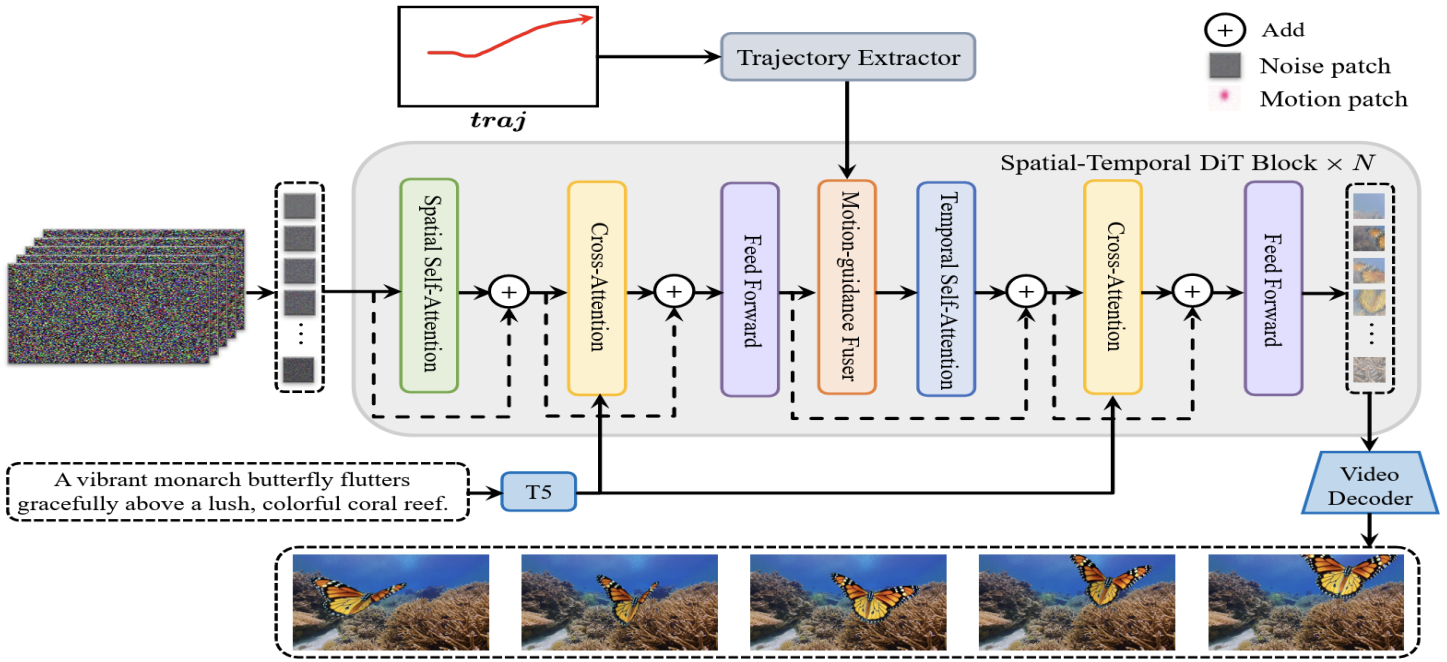


First frame shows the input trajectory for the generated video

**Time Frame**

Tora: Trajectory-oriented diffusion transformer for video generation, CVPR 2025

# But trajectory methods need extra inputs at inference




- Require **explicit trajectory signals** (paths, flow, pose) — not always available at inference
- Extra supervision during training (trajectory data)
- Generalization limited: **motion ≠ physics** (object may follow path but still ignore gravity or collisions)

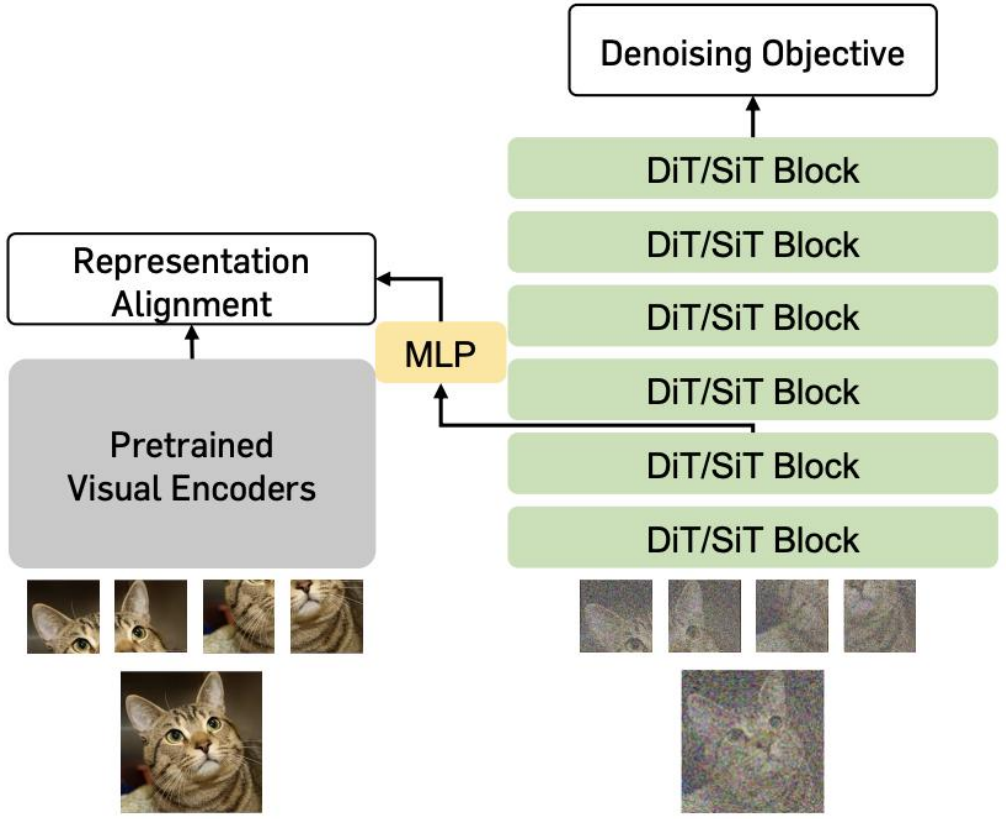
Tora: Trajectory-oriented diffusion transformer for video generation, CVPR 2025

# REPA: Representation Alignment for Diffusion Models

- Trains diffusion models with an **auxiliary alignment loss**
- Aligns **diffusion features** with **teacher network features** (e.g., CLIP, DINO)
- Improves **semantic faithfulness** and **robustness** in generation

$$\mathcal{L}_{\text{REPA}}(\theta, \phi) := -\mathbb{E}_{\mathbf{x}_*, \boldsymbol{\epsilon}, t}\left[\frac{1}{N}\sum_{n=1}^{N} \text{sim}(\mathbf{y}_*^{[n]}, h_\phi(\mathbf{h}_t^{[n]}))\right]$$

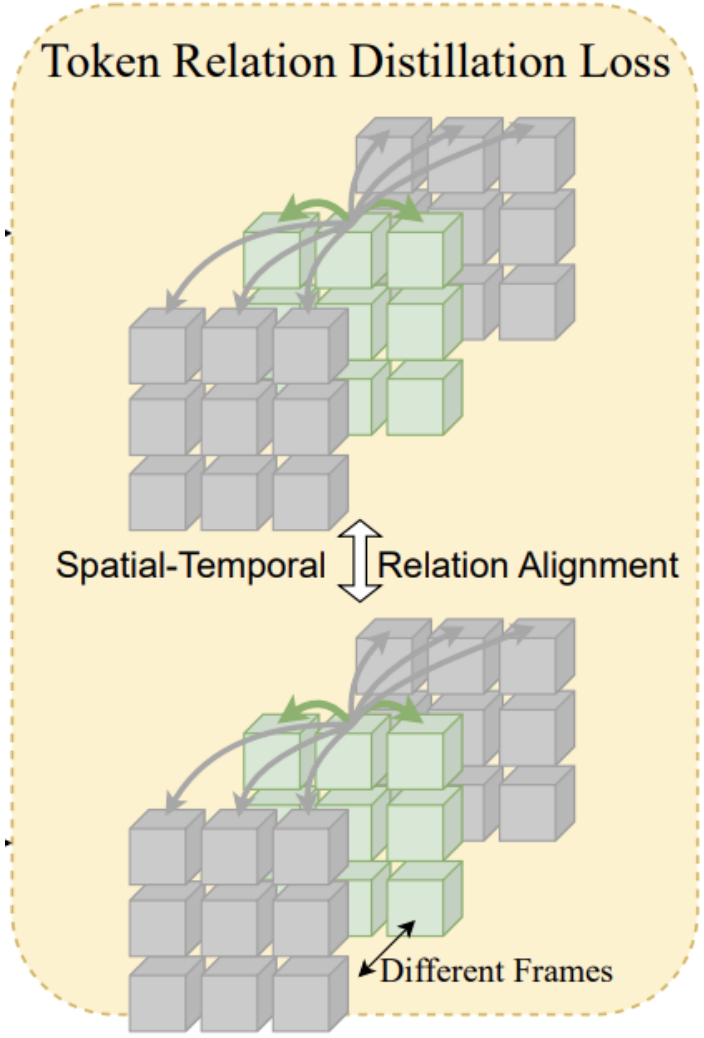
Where $y_*^{[n]}$ : teacher feature, $h_\phi$ : Projector, $h_t^{[n]}$ : diffusion feature




Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think, ICLR 2025 Oral

# VideoREPA aligns diffusion features with motion-aware teachers

- Extends REPA from images to the **video domain**
- Uses a **video teacher** (e.g., VideoMAE) to provide motion-aware representations
- **TRD loss**: aligns **feature relations** across frames, not raw features
- Improves **temporal coherence** and **semantic grounding** in generated videos

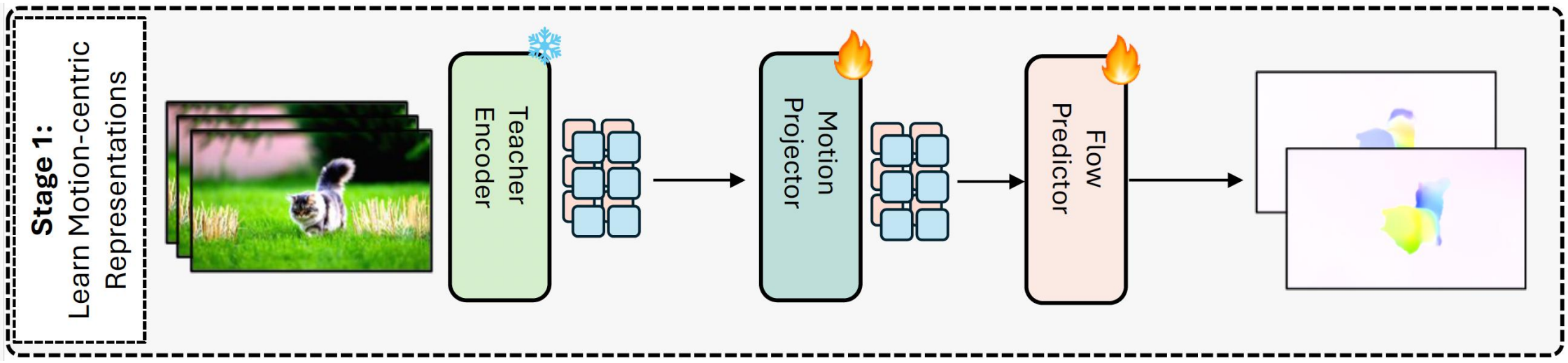



VideoREPA: Learning Physics for Video Generation through Relational Alignment with Foundation Models, Neurips 2025

# VideoREPA aligns with entangled features, not pure motion

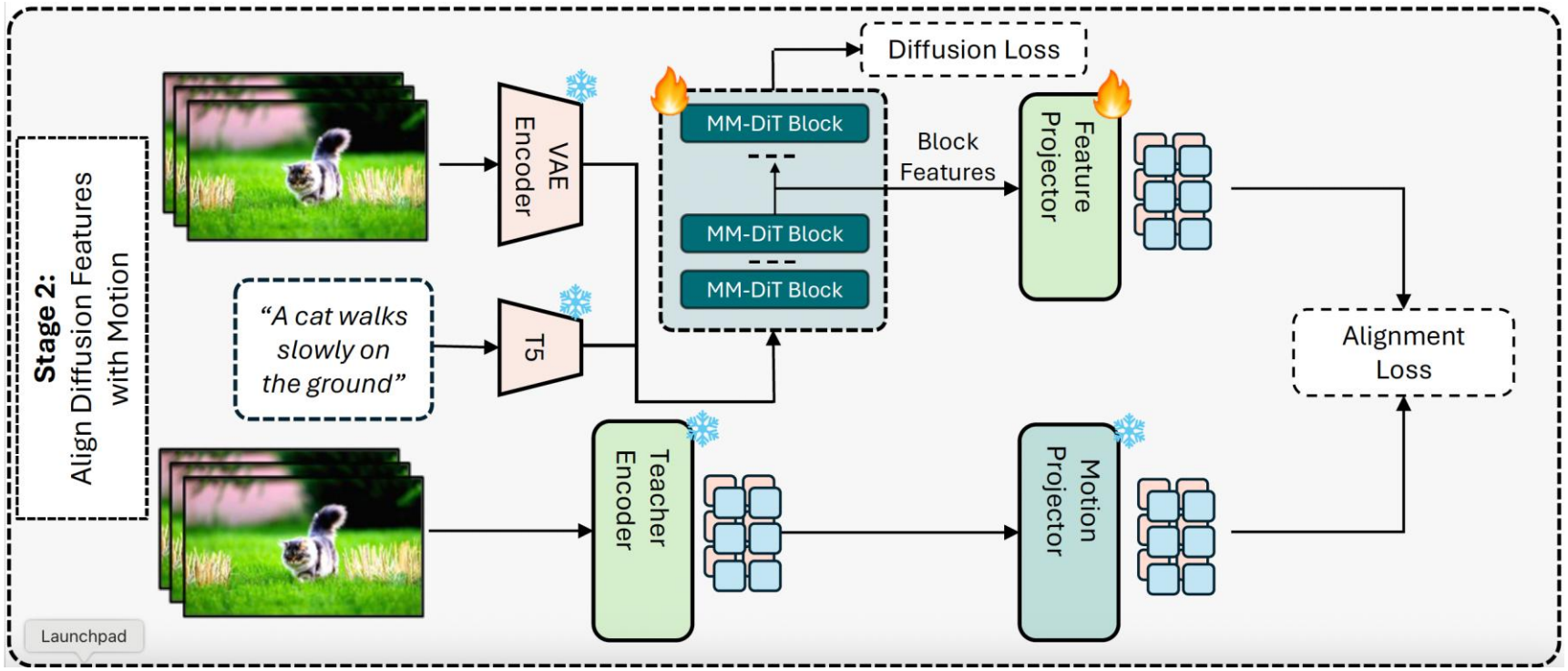
- Teacher features (VideoMAE) are **entangled**: motion + appearance mixed
- Alignment leans toward **appearance**, weak on motion grounding
- Need: align with a **disentangled motion subspace**
- Challenge:
  - Ensuring it's truly **motion-focused**
  - How to **control / supervise** disentanglement

# Stage 1: Learning a disentangled motion subspace




- Start with **VideoMAEv2 features** (teacher encoder)
- Train a **projector** to compress features into a **motion-only space**
- Supervise with **optical flow** to ensure the space captures dynamics
- Result: a disentangled motion subspace, separated from appearance cues

# Stage 2: Aligning diffusion features to motion subspace




- Take **diffusion model features** and project them into the motion space
- Compare them to teacher motion features via a **soft relational loss**
- Encourages diffusion features to respect **motion dynamics & coherence**

# Soft Relational Alignment (our loss)

- Setup (tokens & similarities):

$$\text{Student (projected)}: \; Z \in \mathbb{R}^{F'' \times H'' \times W'' \times D_m}, \quad \text{Teacher (motion)}: \; M \in \mathbb{R}^{F'' \times H'' \times W'' \times D_m}$$

$$\text{Frame } f: \; Z_f, M_f \;\Rightarrow\; Z_f^\flat, M_f^\flat \in \mathbb{R}^{(H''W'') \times D_m}$$

$$\text{Cosine sim: } \; \mathrm{sim}(a,b) = \frac{a^\top b}{\|a\| \, \|b\|}$$

$$S_Z^{\text{spatial}}(f)[i,j] = \mathrm{sim}(Z_{f,i}, Z_{f,j}), \qquad S_M^{\text{spatial}}(f)[i,j] = \mathrm{sim}(M_{f,i}, M_{f,j})$$

$$\text{Flatten tokens over time: } Z^{(i)}, M^{(i)} \Rightarrow S_Z^{\text{temp}}[i,j] = \mathrm{sim}\left(Z^{(i)}, Z^{(j)}\right), \quad S_M^{\text{temp}}[i,j] = \mathrm{sim}\left(M^{(i)}, M^{(j)}\right)$$

$$W_{ij} = \begin{cases} \exp\left(-\dfrac{\Delta_{ij}}{\tau}\right), & \Delta_{ij} \neq 0 \\ 0, & \Delta_{ij} = 0 \end{cases} \quad \text{(exclude intra-frame pairs; } \tau \text{ is temperature)}$$

- Soft relational alignment loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{F''} \sum_{f=1}^{F''} \left\| S_Z^{\text{spatial}}(f) - S_M^{\text{spatial}}(f) \right\|_1 + \left\| W \odot S_Z^{\text{temp}} - W \odot S_M^{\text{temp}} \right\|_1$$

# We evaluate physics, quality, and human preference

- 🧍 **VideoPhy2 (action-centric):** Focuses on human–object interactions; a recent benchmark designed to test **physical plausibility** in action scenarios.
- 🧪 **VideoPhy (material-centric):** Covers **solid–solid, solid–fluid, and fluid–fluid** interactions; complements VideoPhy2 by stressing **materials/interaction types**.
- 🎞️ **VBench and VBench-2.0 (quality toolkit):** Measures **perceptual/technical characteristics** (e.g., aesthetics, temporal smoothness, object–scene consistency).
- 👀 **Blind user study:** Human side-by-side preference test across your models and baselines, mixing prompts from **VideoPhy2** and **VBench-2.0**.

# VideoPhy2 tests physics in action-centric scenes

- **Protocol:** Generate from **591 extended prompts**; scored by **VideoPhy2-AutoEval**.
- Checks **Semantic Adherence (SA)** and **Physical Commonsense (PC)** on a **5-point** scale.
- **Primary focus: Joint** = fraction of videos **≥4 on both** SA and PC.
- a **static baseline** (repeat the first frame) can look "physically safe" and score high on PC

| Method | SA | PC | Joint |
|---|---|---|---|
| CogVideoX-2B | <u>27.1</u> | 64.5 | 22.3 |
| Static baseline | 15.6 | **91.0** | 15.1 |
| CogVideoX-2B (FT) | 26.4 | 73.1 | 22.8 |
| VideoREPA-2B (paper) | 21.0 | 72.5 | – |
| VideoREPA-2B (reimpl.) | 26.1 | 73.3 | <u>23.0</u> |
| MoAlign-2B (ours) | **28.8** | <u>75.0</u> | **24.9** |

# VideoPhy stresses material-centric interactions

- **Protocol: 343 prompts**; evaluated with **VideoConPhysics** auto-rater.
- Emphasizes **materials/interaction types** rather than action scenarios.
- Only **SA** and **PC** scores are computed.

| Method | Solid–Solid | | Solid–Fluid | | Fluid–Fluid | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | SA | PC | SA | PC | SA | PC | SA | PC |
| CogVideoX-2B | **24.7** | 16.9 | **67.5** | 24.8 | **69.0** | 40.0 | **49.8** | 23.9 |
| CogVideoX-2B (FT) | 22.5 | 29.6 | 62.1 | 34.5 | 58.2 | 45.5 | 44.9 | 34.1 |
| VideoREPA-2B (reimpl.) | 23.2 | <u>31.0</u> | <u>66.9</u> | <u>39.3</u> | 54.6 | <u>52.7</u> | 46.7 | <u>37.9</u> |
| MoAlign-2B (ours) | **24.7** | **31.7** | <u>66.9</u> | **40.7** | <u>67.3</u> | **56.4** | <u>49.3</u> | **39.4** |

# VBench(1) checks quality; VBench-2.0 checks faithfulness

- **Purpose:** ensure physics gains **don't degrade overall quality**. Use two complementary toolkits.
- **VBench:** focuses on **perceptual/technical characteristics** — aesthetics, temporal smoothness, object–scene consistency, etc.
- **VBench-2.0:** targets **intrinsic faithfulness** — dimensions like **instance preservation, dynamic spatial relationships, human anatomy,** etc
- Together, they probe **quality vs. faithfulness**, complementing VideoPhy/VideoPhy2's physics focus

| Model | VBench | | | VBench-2.0 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | Quality | Semantic | Total | Creativity | Commonsense | Controllability | Human Fidelity | Physics |
| CogVideoX-2B | 80.6 | 81.6 | 76.6 | 54.9 | 52.8 | 60.2 | **26.6** | 81.1 | **53.9** |
| CogVideoX-2B (FT) | 80.3 | 81.1 | 77.1 | 54.7 | **58.7** | 60.8 | 25.6 | 83.4 | 44.9 |
| VideoREPA-2B | 80.5 | 81.3 | 77.2 | 55.0 | 56.9 | 61.4 | 25.9 | 85.4 | 45.1 |
| MoAlign-2B (ours) | **81.3** | **82.0** | **78.2** | **55.9** | 52.8 | **65.5** | 25.7 | **86.7** | 48.8 |

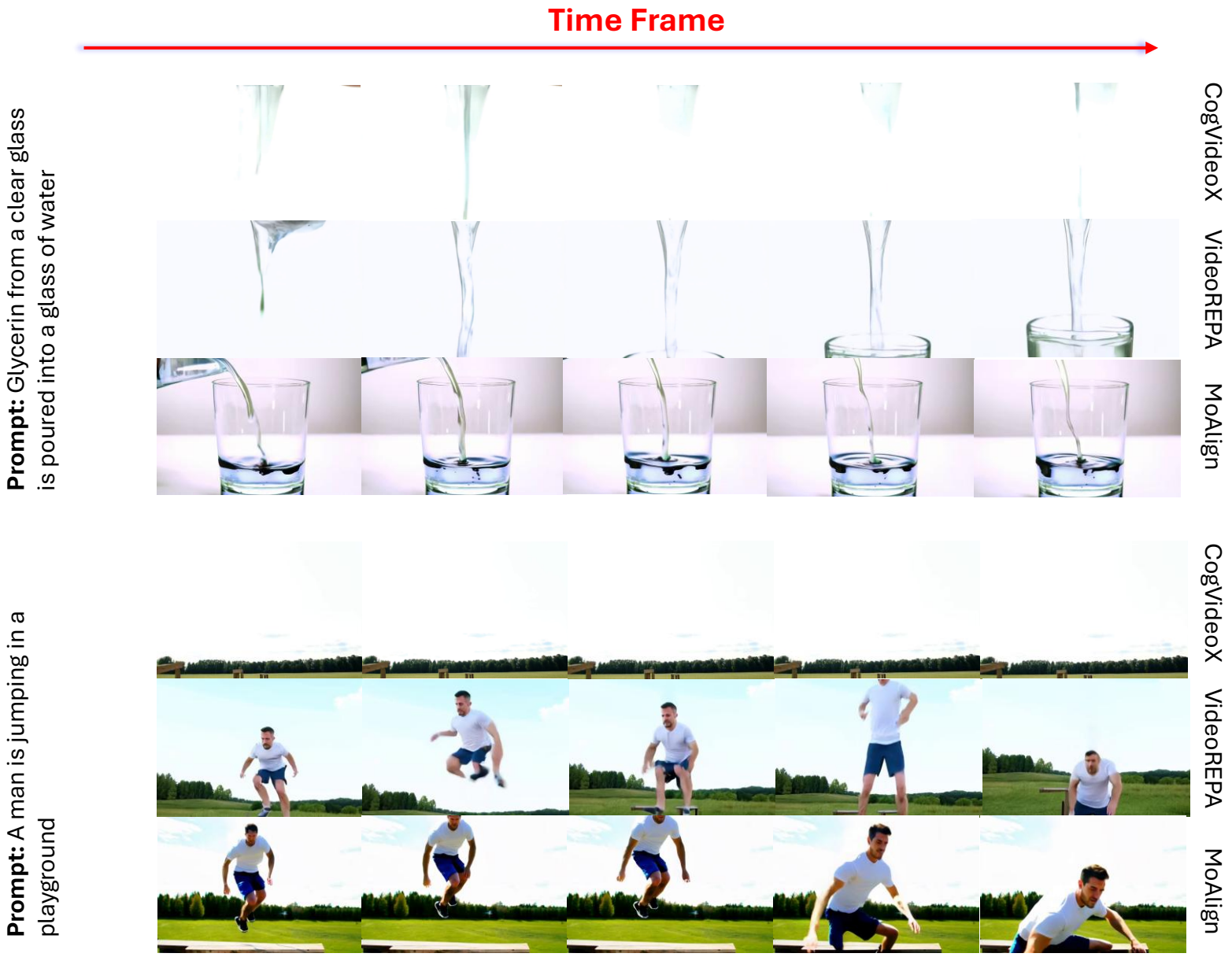
# Blind user study: MoAlign is preferred

**Setup (short):** 3 models (**CogVideoX-2B**, **VideoREPA-2B**, **MoAlign-2B**).

| Comparison | MoAlign | Baseline |
|---|---|---|
| vs. CogVideoX-2B | **68%** | 32% |
| vs. VideoREPA-2B | **78%** | 22% |

**50 videos/model** from a mix of **VBench-2.0 + VideoPhy2** prompts; **672** side-by-side **pairwise** judgments.

# Qualitative Examples



# Limitations: motion can be damped; data + FT matter

- some generations show **reduced motion intensity** (more conservative dynamics)
    - **Data bias** — training set under-represents **strong/varied motion**, so Stage-1 learns a motion space skewed to mild dynamics.
    - **Fine-tuning (FT)** — FT on a **narrow distribution** → **overfitting**; semantics stay, but motion gets **dampened**.
- physics/consistency improve, but **liveliness** of motion can drop on certain prompts.
- **Future Fixes:**
    - **Data curation:** include **motion-rich clips**, balanced motion amplitudes & interaction types, longer spans; avoid static-heavy subsets.
    - **Better alignment (Stage-2):** Be it in terms of a better alignment loss, or some clever FT strategies

# Our motion subspace is improvable

- The learned motion subspace is **not optimal**—captures coarse dynamics but misses fine structure.
- At the moment, no guarantee of complete disentanglement
- Future Fixes:
  - Use long-range **point trajectories / tracklets** as pseudo-labels to shape the subspace.
  - **Task-aware mixture of subspaces:** Learn **specialized projectors** (ballistic, contact/interaction, fluid-like), with a light **router** (prompt/video probe) to pick experts. Have option for semantic separation.

# THANK YOU