STANCE: Controllable Video Generation for Structured Dynamics via Sparse-To-dense ANChored Encoding

Zhifei Chen[1,*] Tianshuo Xu[1,*] Leyi Wu[1] Luozhou Wang[1] Dongyu Yan[1]
Zihan You[3] Wenting Luo[3] Guo Zhang[4] Yingcong Chen[1,2,†]

[1]HKUST(GZ) [2]HKUST [3]XMU [4]MIT

[*] Equal contribution [†] Corresponding author

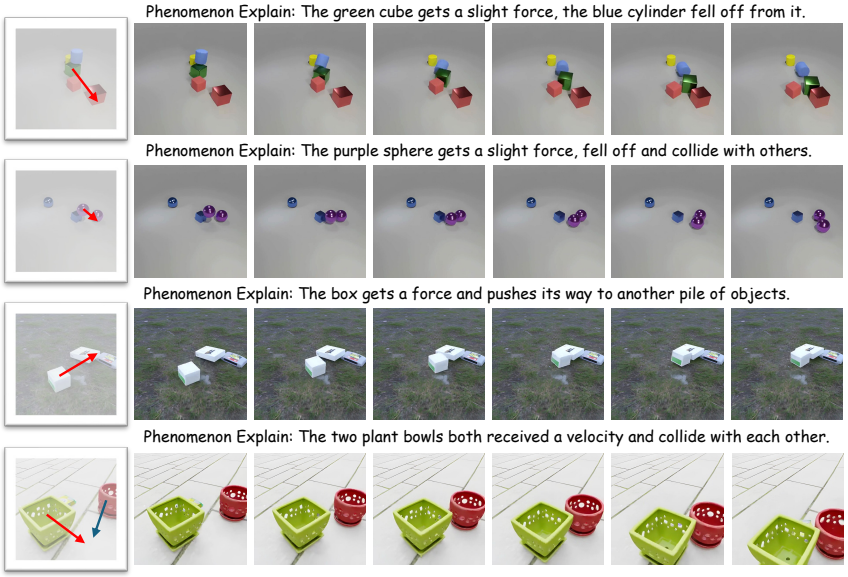

Fig. 1: Videos generated by **STANCE**. User input: one keyframe, coarse 2D arrows, per-instance mass, and a scalar depth delta Δz. Controls yield physically meaningful edits while preserving appearance: increasing mass can reverse collision outcomes, larger speeds produce longer travel and earlier contact, and rotating the arrow reorients trajectories and shifts contact points; Δz disambiguates out-of-plane intent under camera motion. Examples span both *simple collision* setups and *realistic scenes*, including gentle pushes that dislodge or trigger collisions.

Abstract. Video generation has recently made striking visual progress, but maintaining coherent object motion and interactions remains difficult. We trace two practical bottlenecks: (i) human-provided motion hints (e.g., small 2D maps) often collapse to too few effective tokens after encoding, weakening guidance; and (ii) Current architectures inevitably bias toward pixel-level texture reconstruction over complex structural motion, yielding videos that are photorealistic but physically incoherent.

We present **STANCE**, an image-to-video framework that addresses both issues with two simple components. First, we introduce **Instance Cues**, a pixel-aligned control signal that turns sparse, user-editable hints into a dense 2.5D (camera-relative) motion field by averaging per-instance flow and augmenting with monocular depth over the instance mask. This reduces depth ambiguity compared to 2D drag/arrow inputs while remaining easy to user. Second, we preserve the salience of these cues in the latent token space with **Dense RoPE**, which tags a small set of initial-frame motion tokens with spatially-addressable rotary embeddings. Finally, rather than treating appearance and structure as loosely coupled tasks, STANCE integrates an auxiliary stream directly within the same Diffusion Transformer (DiT). By sharing identical latent tokens and Dense RoPE, this auxiliary stream acts as a "geometry witness" that heavily penalizes geometric drift.

1 Introduction

Recent progress in controllable video generation [1, 4, 10, 29, 30] has made it possible to synthesize video clips with rich appearance and diverse dynamics, fueling applications in entertainment, XR, driving, and robotics. Yet, despite impressive visual quality, maintaining logical and physical coherence like consistent object trajectories, inertia-like motion, and interaction plausibility remains challenging for large video diffusion/autoregressive backbones. Despite strong recent progress, ensuring temporal and interaction coherence remains non-trivial. In practice, models that excel at appearance can still exhibit small motion inconsistencies like drift in trajectories or ambiguous contacts, especially when guidance comes from simple control inputs.

One line of work [15] tackles coherence by conditioning on full trajectories, which can stabilize object motion but assumes frame-level scripts or dense supervision. In practice, this is rarely available or easy to edit. We focus on a concrete, widely relevant regime: rigid object interactions with contacts, where plausibility depends on getting interaction timing and motion continuity right. These aspects are easy for humans to specify coarsely (e.g., directions, speed hints, relative sizes) but hard to author frame by frame. Rather than replacing the generative prior with a trajectory controller, we keep the prior in the loop and steer it using sparse, human-editable cues that are lifted into a dense, model-friendly representation.

From a modeling perspective, incoherence does not stem solely from missing "physics." Two pragmatic factors often erode controllability: (i) sparse, low-resolution control maps (especially when injected at a single time slice), can be washed out by tokenization and early attention, leaving too few effective tokens to guide the backbone; (ii) objectives that couple appearance and motion can induce trade-offs, where improving visual quality often comes at the expense of motion consistency. Therefore, a useful control pathway should remain token-dense after encoding, preserve precise spatial alignment, and enable high-quality synthesis while maintaining coherent motion. In our setting, control tokens en-

code initial conditions (who moves, where, how fast, with what mass) rather than full future trajectories, and the DiT is responsible for learning physically reasonable rollouts of rigid-body collisions.

Concretely, we introduce **"Instance Cues"** as a pixel-aligned motion control. During training, we derive per-instance average flow (augmented with monocular depth) and spread it over the instance mask. Unlike 2D drag/arrow interfaces [19, 33] which lack depth awareness and can be ambiguous under camera motion, our depth-augmented cues encode a direction with depth (2.5D, camera-relative), improving spatial disambiguation. We further introduce a **"Dense RoPE"** mechanism: instead of passing a single low-res map, we select salient spatial locations and assign high-salience, spatial-addressable rotary embeddings to the corresponding motion tokens. Then, we jointly synthesize RGB and a structural stream (segmentation or depth) under the same instance cues. The two streams share spatio-temporal tokens and attend to the same cue tokens, so the structural head regularizes the RGB head, tightening alignment and reducing drift without requiring per-frame scripts. In summary, we make three main contributions:

- **Pixel-aligned, human-editable control.** We convert sparse user hints into a dense, pixel-aligned *2.5D* motion field. Compared with pure 2D drag or arrow inputs, our depth-augmented cues disambiguate motion under camera movement, remain easy to author, and preserve appearance while steering direction, speed, and contact outcomes across multiple objects.
- **Token-dense injection via Dense RoPE with a structural witness.** To keep control effective after encoding, we select nonzero sites in each target region and assign a fixed budget of motion tokens tagged with *first-frame* RoPE, preserving spatial identity over time. We jointly train RGB with a lightweight structural head (depth) that attends to the same cues, acting as a geometry witness.
- **Comprehensive data and validation.** We curate a 200k-clip dataset of rigid-object interactions spanning single and multi-object cases as well as realistic scenes, and run extensive ablations isolate the contributions of Dense RoPE and the auxiliary structural stream, showing gains in control faithfulness (direction, speed and mass), temporal coherence (reduced hover and drift), and interaction plausibility (cleaner contact onsets).

2 Related Works

2.1 Video Diffusion Models

Recently, diffusion models have demonstrated remarkable progress across a variety of video generation tasks. The ability to effectively capture complex spatio-temporal dependencies, making diffusion models well-suited for video generation where both high visual quality and frame-to-frame consistency are essential. Early video diffusion models [1, 9, 26] were primarily built on UNet-based frameworks, which has a symmetric encoder–decoder structure with skip connections.

This design facilitates efficient extraction of spatial information, while temporal attention modules are often incorporated to further enhance temporal consistency across frames [8].

More recently, Diffusion Transformers (DiTs) have become the dominant architecture for video generation [13, 16, 25, 29, 32]. By leveraging self-attention mechanisms, DiTs achieve superior modeling of long-range spatio-temporal dependencies, leading to significant improvements in both visual quality and temporal coherence. Compared with UNet-based architectures, DiTs provide greater flexibility in scaling to large models and datasets, better parallelization during training and inference, and a unified architecture that naturally accommodates multi-modal conditioning.

2.2 Motion-conditioned Video Generation

While recent video diffusion models show impressive visual quality, maintaining coherent motion remains challenging. Unlike pure text-to-video generation—where motion is implicitly induced by abstract prompts—motion-conditioned video generation must accept *explicit* signals that steer dynamics and better align with user intent. Flow-based conditioning has been explored to inject motion cues into the generative process [5, 19, 20], and drag-based interfaces let users specify trajectories by placing start/end handles on object parts [14, 27, 30].

Despite this progress, important gaps persist. First, many control signals are either cumbersome to author or too sparse after encoding to effectively shape dynamics; downsampling and tokenization can wash out weak cues, especially for small or thin objects. Second, temporal coherence can break around contacts: models may hover before impact, mis-time contact onsets, or exhibit bounce-backs without contact, revealing a mismatch between appearance fidelity and interaction plausibility. Third, most methods cannot modify object properties (e.g., mass), which limits faithfulness to user intent and basic physical expectations. For instance, in a collision between a large ball and a smaller one, a user might specify a heavier small ball that should dominate the interaction; in practice, model priors often enforce the opposite outcome. Concurrently, VACE [11] proposes an all-in-one DiT-based framework for unified video creation and editing, organizing diverse multimodal inputs (text, frames, masks) into a Video Condition Unit (VCU) and injecting them via a Context Adapter. VACE focuses on broad task coverage, but does not target object-level drag-based control with explicit supervision. Our Instance Cues, Dense RoPE, and joint auxiliary head are complementary to this line of work, aiming instead at a learned, physics-aligned rigid-body simulator under controllable motion editing.

3 Method

3.1 Preliminary

Recent open-source systems adopt diffusion transformers (DiT) for video generation [22, 29]. Two design shifts distinguish these models from earlier approaches:

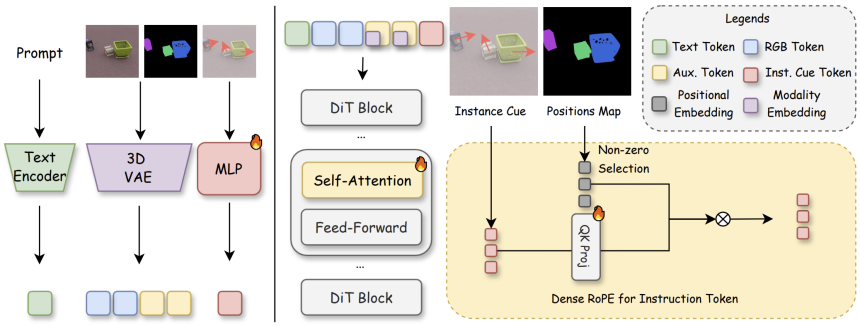

Fig. 2: Pipeline of STANCE. Our method is organized as follows: (1) **Left:** we extend the input of DiT to include new alpha tokens, and use a train-able MLP to tokenize instance cues. (2) **Right:** The modality embeddings are added to the auxiliary tokens, and the instance cue tokens are paired with Dense RoPE.

(i) instead of alternating 1D temporal and 2D spatial attention blocks [2–4, 32], they apply a single 3D spatio-temporal self-attention; *(ii)* text tokens are concatenated with visual tokens and the entire sequence is processed by full self-attention (rather than text-only cross-attention). Full self-attention is then applied across the combined sequence:

$$\begin{aligned} &\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad \text{where} \\ &\mathbf{Z} : \mathbf{Z} \in \{\mathbf{Q},\mathbf{K},\mathbf{V}\} \\ &= [\mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{text}}); \mathbf{f}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{video}})] \end{aligned} \tag{1}$$

Here $\mathbf{W}_t$ (for $t \in \{q,k,v\}$) represents the projection matrixs in the transformer model, and $\mathbf{f}_t$ (for $t \in \{q,k,v\}$) represents a combined operation that incorporates both the projection and positional encoding for visual tokens. A key modeling choice is the positional encoding for video tokens (indexed by a spatio-temporal position m) prior to projection.

There are two commonly used types of positional encoding. One is absolute positional encoding formulated as follows:

$$\mathbf{f}_{z:z\in\{q,k,v\}}(\mathbf{x}_{\text{video}}) := \mathbf{W}_{z:z\in\{q,k,v\}}(\mathbf{x}^m_{\text{video}} + \mathbf{p}^m), \tag{2}$$

where $\mathbf{p}$ is the positional embedding (e.g., a sinusoidal function) and m denotes the position of each RGB video token. Another approach is the Rotary Position Embedding (RoPE) [21], often used by [22, 29]. This is expressed as

$$\mathbf{f}_{z:z\in\{q,k\}}(\mathbf{x}_{\text{video}}) := \mathbf{W}_{z:z\in\{q,k\}}(\mathbf{x}^m_{\text{video}}) \circ e^{im\theta}, \tag{3}$$

where m is the positional index, i is the imaginary unit for rotation, and θ is the rotation angle.

3.2 Our Approach

Figure 2 illustrates STANCE's pipeline. Given a text prompt and a keyframe, the user supplies instance masks, coarse arrows, and a mass tag per instance. We convert these sparse inputs into a dense, pixel-aligned instance cue field (2.5D, camera-relative) and inject it into the model with Dense RoPE tagging. The model jointly predicts RGB and an auxiliary structural map (segmentation or depth), sharing spatio-temporal tokens and attending to the same cue tokens.

Sparse→Dense Motion Cues We use *instance cues* as the control signal: a few per–object motion hints that are expanded into a dense, mask–aligned field (2.5D when depth is used). t remains easy to use.

Training. For each clip we have optical flow $\mathbf{O}$ between two reference frames and an instance map on the first frame. For every instance i with pixel set $\Omega^{(i)}$, we compute a *mean motion vector* by averaging flow over the instance, then *paint* this vector back over the renderer-provided ground truth mask to obtain a dense field:

$$\bar{\mathbf{v}}^{(i)} \;=\; \frac{1}{|\Omega^{(i)}|} \sum_{(x,y)\in\Omega^{(i)}} \mathbf{O}(x,y)\,.$$

With monocular depth provided, analogous to optical flow, we derive a per-instance *delta depth*: given monocular depths D_t and D_{t+1}, we set $\Delta z_i = \mathrm{mean}_{\mathbf{p}\in M_i}\big(D_{t+1}(\mathbf{p}) - D_t(\mathbf{p})\big)$, and append this scalar as the third control channel, yielding a camera-relative "2.5D" cue.

Inference. A user specifies a keyframe, per-instance masks (e.g., SAM [12]), and a coarse 2D arrow for each instance, together with a mass value. We rasterize each arrow inside its mask to obtain a dense in-plane control map. Optionally, the user provides a scalar depth delta Δz per arrow, which we broadcast over the same mask and *append as a third control channel* (as in training) to disambiguate motion under camera movement. The per-instance mass is fused in exactly the same manner as Δz: the scalar is broadcast over the instance mask and appended as a *fourth* channel of the dense instance-cue field (flow u, v, Δz, mass m) before patchify, so it stays spatially aligned and is naturally selected by Dense RoPE alongside the motion channels. In Fig. 3, the vertical axis depicts the user-drawn in-plane arrow ($\Delta z = 0$, black); a positive depth delta (red, right of 0) indicates motion into the screen, whereas a negative depth delta (blue, left of 0) indicates motion out of the screen.

3.3 Dense RoPE

Motivation: Downsampling during *patchify* often makes the 2D control mask extremely sparse: a few informative value are surrounded by many zeros, particularly for small or thin region. We keep only the nonzero sites and enforce a fixed token budget, guaranteeing enough motion tokens even for tiny objects.

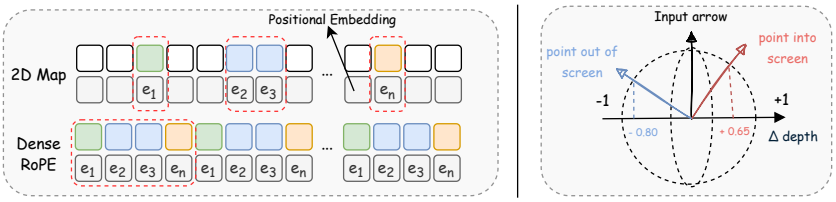

Fig. 3: Left: *2D Map vs. Dense RoPE.* When the 2D control map is downsampled, many tokens inside the window become zeros, yielding a sparse signal that weakens control. **Dense RoPE** performs non-zero token extraction over the target region (colored), preserves/assigns positional embeddings $(e_1, \ldots, e_n)$, and feeds a compact, dense sequence to the model—resulting in stronger, spatially focused control. **Right:** *Depth control.* The upward black arrow is the user-drawn 2D arrow; by manipulating a scalar depth delta (Δ depth on the horizontal axis, $[-1, +1]$), the user specifies out-of-plane motion: $\Delta > 0$ (red) points *into* the screen (away from the camera), while $\Delta < 0$ (blue) points *out of* the screen (toward the camera).

Each selected token is tagged with its first-frame RoPE so its spatial identity persists over time; these tokens act as stable, high-signal anchors that later layers can reliably attend to. Unlike global rescaling, this directly reduces dilution and keeps the control pathway token-dense and spatially aligned after encoding.

Let the first–frame control mask be $\mathbf{M} \in \{0,1\}^L$ on the latent token grid, and let $\mathbf{X}_{\text{Cue}} \in \mathbb{R}^{L \times C}$ denote the per-token control features (e.g., patchified 2.5D instance-cue latents). We collect active indices

$$\Omega \;=\; \{\, i \in \{1, \ldots, L\} :\; \mathbf{M}_i = 1 \,\}, \qquad m = |\Omega|.$$

To meet a fixed budget N of motion tokens required by the backbone, we form an index list $\mathcal{J}$ by

$$\mathcal{J} \;=\; \begin{cases} \text{uniformly subsample } \Omega \text{ to length } N, & m > N, \\ \text{tile and truncate } \Omega \text{ to length } N, & m \le N, \end{cases} \quad \Rightarrow \quad |\mathcal{J}| = N.$$

Given the selected index set $\mathcal{J}$ and per-token flow features $\mathbf{x}_j^{\text{Cue}} \in \mathbb{R}^C$ for $j \in \mathcal{J}$, we form query $\mathbf{q}$, key $\mathbf{k}$, for the motion tokens using the same $\mathbf{f}_z$ operator as visual tokens, where $\mathbf{p}^j$ is the *first-frame* positional code at site j, and scaled by a learnable gain g:

$$\mathbf{q}_j^{\text{Cue}} \;=\; \mathbf{f}_q\big(\mathbf{x}_j^{\text{Cue}}\big) = \mathbf{W}_q\big(\mathbf{x}_j^{\text{Cue}} + \mathbf{p}_j^{\text{Cue}}\big), \quad \tilde{\mathbf{k}}_j^{\text{Cue}} \;=\; g_k\, \mathbf{f}_k\big(\mathbf{x}_j^{\text{Cue}}\big) = g_k\, \mathbf{W}_k\big(\mathbf{x}_j^{\text{Cue}} + \mathbf{p}_j^{\text{Cue}}\big), \tag{4}$$

we then concatenate motion tokens into the full sequence. The detailed algorithm flow can be located in the supplementary section.

3.4 Joint Auxiliary Generation

Joint RGB + auxiliary map generation. We extend the pretrained RGB video backbone to jointly synthesize an *auxiliary* structural stream (segmentation or

depth) alongside RGB. Concretely, we duplicate the video token sequence so the model handles two modality-aligned streams of equal length L: the first L tokens decode to RGB and the next L tokens decode to the auxiliary map:

$$\mathbf{X}^{1:2L}_{\text{video}} \;=\; [\,\mathbf{X}^{1:L}_{\text{rgb}};\; \mathbf{X}^{1:L}_{\text{aux}}\,].$$

Positional alignment with a light domain tag. RGB and auxiliary tokens at the *same* spatio–temporal index share the *same* positional code to enforce pixel/time alignment; a tiny learnable domain embedding marks the additional modality.

$$\mathbf{f}_z(\mathbf{x}^{\text{video}}) \;=\; \mathbf{W}_z\big(\mathbf{x}^{\text{video}} + \mathbf{p}^{\text{video}}\big), \qquad \mathbf{f}^*_z(\mathbf{x}^{\text{aux}}) \;=\; \mathbf{W}^*_z\big(\mathbf{x}^{\text{aux}} + \mathbf{p}^{\text{aux}} + \mathbf{d}_{\text{aux}}\big), \tag{5}$$

where $\mathbf{d}_{\text{aux}}$ is a learnable, zero-initialized domain vector; for RoPE we simply apply the *same* rotary index θ_m to both RGB and auxiliary keys/queries at m.

Self-attention is applied over text and both streams:

$$\mathbf{Z} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} \;=\; \big[\,\mathbf{W}_z(\mathbf{x}_{\text{text}})\;;\; \mathbf{f}_z(\mathbf{x}^{1:L}_{\text{rgb}})\;;\; \mathbf{f}^*_z(\mathbf{x}^{1:L}_{\text{aux}})\,\big]. \tag{6}$$

Training objective. We keep the diffusion objective and supervise both heads; the joint loss is

$$\mathcal{L} \;=\; \mathbb{E}_{t,\epsilon}\Big[\big\|\hat{\epsilon}_{\text{rgb}} - \epsilon\big\|_2^2 \;+\; \lambda_{\text{aux}}\big\|\hat{\epsilon}_{\text{aux}} - \epsilon\big\|_2^2\Big], \tag{7}$$

with a single weighting λ_{aux}. Joint prediction stabilizes optimization: the auxiliary stream anchors structure/geometry while RGB focuses on appearance. Empirical comparisons are reported in Sec. 4.2.

4 Experiments

Dataset Preparation. We build our dataset on Kubric [7], rendering 200k short clips of rigid-body interactions split evenly between (i) simple scenes with single- or multi-ball interactions and (ii) composite realistic scenes with scanned objects and randomized backgrounds. In the simple subset we place one or more rigid objects in a minimal environment and randomize object shape (ball vs. a small set of primitives), mass, initial linear velocity, and initial position/orientation; lighting uses three rectangular area lights plus a directional sun with fixed placements and randomized intensity/temperature. In the composite subset we replace primitives with GSO [6] assets and render against backgrounds sampled from 5,000 environment maps, randomizing object selection, placement, and pose to induce diverse contacts and occlusions. Camera intrinsics/extrinsics and renderer settings are kept consistent within each scene, while material, friction, restitution, and object counts are sampled within bounded ranges to diversify collisions. We construct held-out validation and test splits from both subsets to avoid scene- or asset-level leakage.

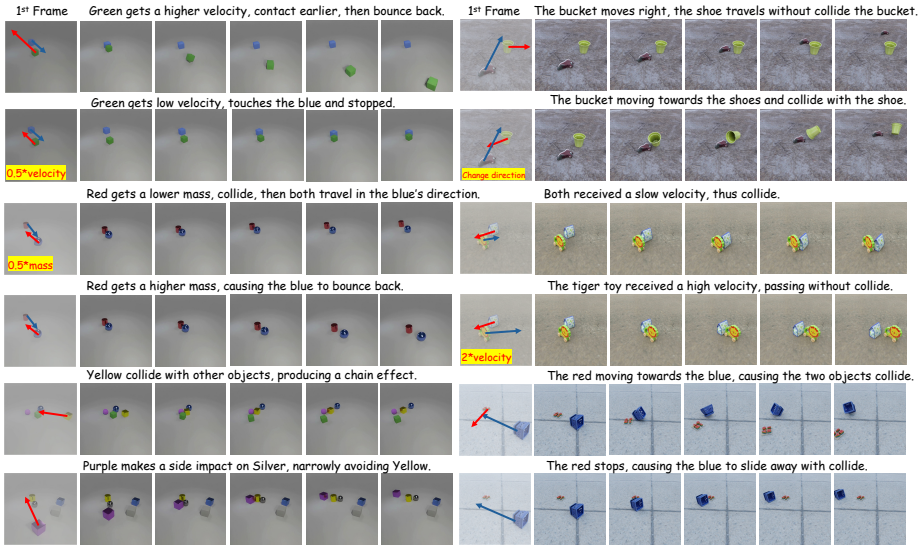

Fig. 4: Applications. Left: Provide with the first frame, we alter the magnitude of the **velocity** and **mass** for the synthetic objects. **Right**: For realistic objects, the mass is fixed, thus we altered the direction and magnitude of velocity only. **The video clips for all shown cases can be located in the HTML files.**

Model. We fine-tune the **CogVideoX-1.5 (5B) Image-to-Video** backbone [29] with our Instance-Cue injection and Dense RoPE. Unless otherwise noted, the model generates **RGB** videos at 512×512 resolution, 49 frames, 16 FPS. We finetune for 50k iterations on 8× H100 GPUs; the base tokenizer and text encoder remain frozen. For domain conditioning, we use a learnable $1 \times D$ vector (zero-init) that is tiled to $L \times D$ per sequence during training. *Matched compute.* The auxiliary stream duplicates only the video tokens; the backbone, attention heads, and parameter count are identical to the RGB-only model, so the gains stem from the control design rather than added capacity. Generating a 49-frame 512^2 clip on a single A100 takes ∼80 s (RGB-only) versus ∼120 s (joint).

Applications. Our interface takes a keyframe with instance masks, per–instance arrows that encode the initial velocity $\mathbf{v}_0$, a scalar depth delta Δz per arrow, and a per–instance mass scalar m. These *Instance Cues* are encoded as our 2.5D (camera–relative) control and injected into the video backbone.

Speed sweep. With mass fixed, varying both the magnitude and direction of the initial velocity $\mathbf{v}_0$ yields predictable kinematics: increasing $\|\mathbf{v}_0\|$ increases displacement over a fixed horizon and reduces time-to-contact, while rotating $\mathbf{v}_0/\|\mathbf{v}_0\|$ reorients the trajectory and shifts the contact point; visual appearance remains unchanged. (Fig. 4, top).

Mass sweep. With $\mathbf{v}_0$ fixed, changing m alters post–contact behavior: mass variation of the red cylinder reverses the collision outcome—when light, it is deflected by the blue ball; after increasing its mass, it instead pushes through and ejects the ball.(Fig. 4, center).

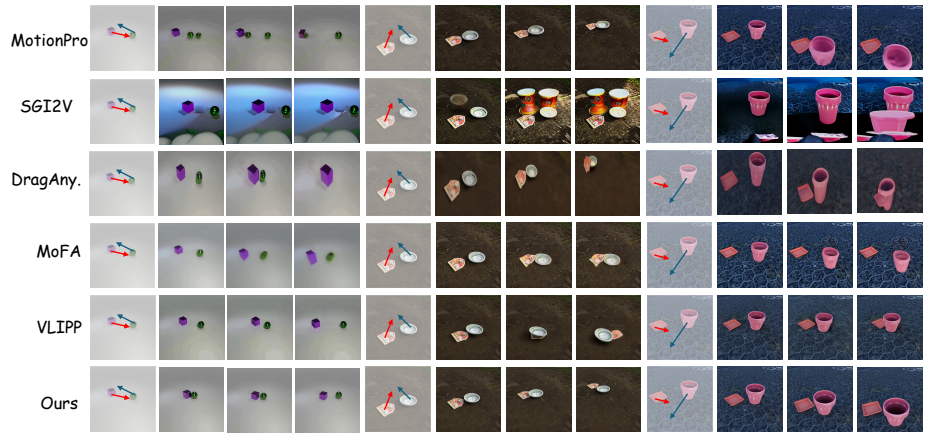

Fig. 5: Qualitative Comparisons. We compare against recent baselines. All videos are temporally aligned (trimmed or padded to a fixed duration) and spatially normalized (resized for visualization). STANCE attains superior visual fidelity while maintaining strong physical coherence. **The video clips for all shown cases can be located in the HTML files.**

In multi–object scenes, editing a single arrow or mass produces coherent chain reactions without frame–level scripts (Fig. 4, bottom-left). Thanks to depth–augmented cues and Dense RoPE, guidance remains spatially localized and temporally consistent.

Evaluation set. We render an additional **200** held-out clips (100 simple, 100 composite) never seen during training. Evaluation is conducted on two splits: *Regular* scenes and a harder *Small-object* subset (thin/tiny instances). Unless otherwise noted, all methods generate 49 frames at 16FPS and 512×512 resolution.

Metric: Physics IQ (↑). To assess motion coherence, we report **Physics IQ** [17]—a single score that emphasizes *how* things move, not just how they look. Physics IQ aggregates a few simple, normalized checks: (i) *Spatial IoU* (where action happens), (ii) *Spatiotemporal IoU* (where and when action happens), (iii) *Weighted Spatial IoU* (where and how much action happens, weighting by motion magnitude), and (iv) *MSE* (how action happens, penalizing deviation from target motion signals). Scores are combined and rescaled to a 0–100 index (higher is better). Compared to common video metrics (e.g., FVD [23]/LPIPS), Physics IQ is more indicative of motion coherence and interaction plausibility, which are the focus of our setting.

4.1 Comparisons

We compare **STANCE** with strong video baselines and editing-by-control methods, and Fig 5 illustrates the outcomes.

Table 1: Ablations on control and joint supervision. Evaluated on a held-out set with *Regular* scenes and a harder *Small-object* subset (thin/tiny instances). Rows compare: (i) *text-conditioned* baseline, (ii) *Instance Cues as a low-res 2D map*, (iii) our *Dense RoPE* motion tokens, and (iv) *joint* RGB+Depth/Seg supervision.

	Physics-IQ ↑		FVD ↓	
	Regular	Small	Regular	Small
Control Signals				
text-conditioned	24.08	—	97.40	—
w/ 2D-Map	<u>43.72</u>	<u>31.92</u>	<u>56.20</u>	<u>58.59</u>
w/ Dense RoPE	**46.89**	**41.83**	**54.63**	**56.32**
Joint Aux. Gen				
Only RGB	46.89	41.83	54.63	56.32
w/ Depth	**49.03**	**45.63**	**50.39**	**51.32**
w/ Segmentation	<u>47.96</u>	<u>45.12</u>	<u>53.09</u>	<u>53.35</u>

Table 2: Baseline comparison. We included a 2B model for fair comparisons. We report *Physics-IQ* (↑; motion coherence/contact plausibility) and *FVD* (↓; perceptual realism) averaged over a 200-clip held-out set. Best and second-best are highlighted.

Method	Physics-IQ ↑	FVD ↓
Baselines		
SG-I2V	15.42	113.54
Drag-Anything	24.86	92.78
MoFA-Video	29.71	98.30
MotionPro	31.58	74.27
VLIPP	36.40	57.90
Ours		
STANCE-2B	45.90	54.97
STANCE-5B	**47.62**	**50.74**

Control faithfulness. Given the same keyframe, instance masks, and arrows, **STANCE** adheres to the intended directions and relative magnitudes (speed/mass edits) while preserving object identity across frames. *VLIPP* [28] specify behavior via prompts rather than pixel-aligned cues; this leaves spatial and metric ambiguities (where, how far, how fast), yielding less precise control than STANCE (Fig. 5).

Contact timing and continuity. STANCE produces cleaner contact onsets and fewer "hovering" frames before/after impact. While *VLIPP* [28] often achieves strong appearance control, its physical consistency is weaker; e.g., in the synthetic case (Fig. 5, left), two objects begin to bounce back before contact. Conversely, *MOFA* [19], *MotionPro* [31] provide precise per-object targeting, but tend to struggle with appearance consistency over time (identity/texture drift and mask leakage under longer sequences or viewpoint changes), whereas our joint RGB+structure training under shared instance cues mitigates these failures.

Quantitative results. We report comparisons in Table 2. Across the 200-clip held-out set, our method attains the highest *Physics IQ* (↑), outperforming *SG-I2V* [18], *Drag-Anything* [27], *MoFA-Video* [19], *MotionPro* [31], and *VLIPP* [28].

Comparison with trajectory-based control. We additionally compare against Wan-Move, a trajectory-conditioned baseline built on a larger 14B backbone (Tab. 3). Despite using far fewer parameters, STANCE-5B matches its Physics-IQ (47.62 vs. 47.50); the heavier 14B backbone attains a lower FVD, as expected from scale. Conceptually, trajectory methods prescribe *how* tokens follow an ex-

Table 3: Comparison with a trajectory-based baseline. STANCE-5B matches a 14B trajectory-conditioned model on Physics-IQ at a fraction of the parameters.

Method	Physics-IQ ↑	FVD ↓
STANCE-5B	**47.62**	50.74
Wan-Move-14B	47.50	**45.78**

Table 4: Disentangled Dense RoPE ablation (small/thin and regular objects merged). Re-samp.: active-site token selection; FF-RoPE: first-frame positional anchoring.

Re-samp.	FF-RoPE	Physics-IQ ↑	FVD ↓
×	×	31.92	66.20
✓	×	*37.45*	*58.18*
×	✓	*35.21*	*54.74*
✓	✓	**37.83**	**54.63**

ternally specified path, whereas STANCE specifies *what* to allocate and lets the DiT roll out its own dynamics.

4.2 Ablation Studies

Dense RoPE. We keep the backbone, training budget identical and compare: (i) text-only CogVideoX, (ii) CogVideoX + 2D–arrow control (single low-res map), and (iii) ours *with* Dense RoPE. On the full held-out set, Dense RoPE consistently improves *Physics IQ* over all baselines. On the ***small-object*** subset, gains are most pronounced: when objects occupy few pixels, the 2D map and naïve tokenization collapse to very few effective tokens, leading to weak guidance; tagging a small set of motion tokens with Dense RoPE preserves spatial addressability and reduces drift/identity swaps under occlusion.
Disentangling Dense RoPE. We further factorize Dense RoPE into its two components—active-site re-sampling and first-frame positional anchoring (FF-RoPE)—and find their contributions additive (Tab. 4). Re-sampling alone alleviates token dilution on small/thin objects, while first-frame anchoring gives each token a stable spatial identity across frames; enabling both yields the best Physics-IQ and FVD.

Auxiliary Ablation To comprehensively validate our joint generation strategy, we evaluate two types of auxiliary modalities and three structural designs for integrating them. These six joint configurations are compared against a standard *RGB-only* baseline.

First, we evaluate how to best incorporate these targets into the DiT architecture (see Fig. 6): (i) *Batch Extension*: Expanding the batch dimension B and introducing a new module for inter-batch communication (similar to [24]). Empirically, this design struggles to capture tight spatial correspondence, resulting

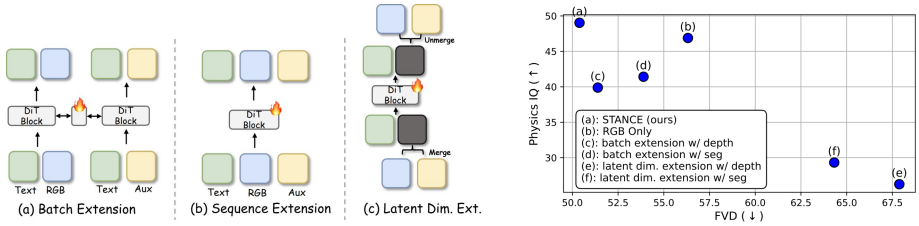

Fig. 6: Ablation of target injection into DiT. Batch extension yields weak control due to poor spatial correspondence; latent-dimension extension degrades diversity by disrupting the pre-trained feature space. Our sequence-level extension provides stronger control and better physical coherence while preserving generative quality.

in weaker controls. (ii) *Latent Dimension Extension*: Merging video and auxiliary tokens along the latent dimension D, projecting them into the DiT's latent space via learnable linear layers, and unmerging them post-generation. While straightforward, this alters the pre-trained feature space, severely degrading generative diversity. Our sequence-level extension achieves the optimal balance (Fig. 6): it facilitates better controls and physical coherence while preserving the generative quality of the pre-trained prior (low FVD).

Next, holding the architecture fixed, we compare the modalities in Tab. 1. Both joint variants (*RGB+Seg* and *RGB+Depth*) improve Physics-IQ over the *RGB-only* baseline, demonstrating that predicting a structural target stabilizes optimization. On the Regular split, depth yields the highest Physics-IQ: its continuous 2.5D cue improves perspective and order reasoning, making contact and motion more coherent than using masks alone. Conversely, on the Small-object split, the performance gap shrinks; segmentation's sharp boundaries provide stronger spatial anchors for tiny or thin targets, whereas monocular depth tends to be noisy at small scales.

4.3 User Study

To further support our editing quality and physical plausibility, we include a user study, and the specifications are listed below: A total of 30 videos are generated, and 87 users participated in the evaluation. The number indicates the users preference in (%). We excluded SG-I2V [18] and MOFA-video [19] due to poor quality.

Table 5: User Study: comparison of physical plausibility and control faithfulness.

Method	Physical plausibility	Control faithfulness
MotionPro	8.7%	11.7%
VLIPP	8.0%	16.0%
STANCE (ours)	**83.3%**	**72.3%**

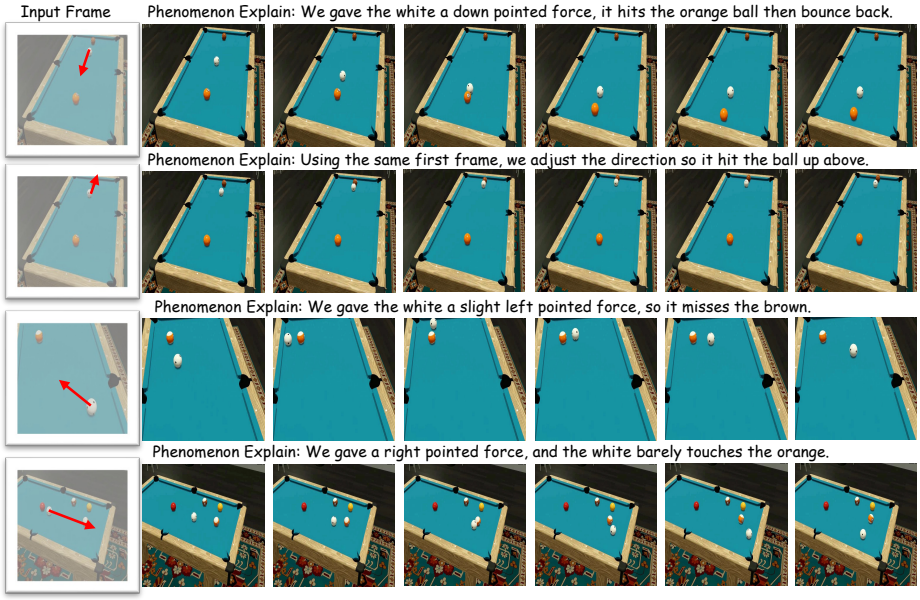

Fig. 7: Pool playing demos. We evaluate **STANCE** by "playing" a pool game. Given the first frame and user-specified initial velocities, STANCE follows the directions/speeds, preserves object identity, and produces physically coherent outcomes.

4.4 Real-World Demonstration

Pool Playing tests. We evaluate STANCE to the playing pool domain (Fig. 7). In this scenarios, the model accurately simulates the elastic collisions between the cue ball and the rack, demonstrating precise momentum transfer.

Real-world captured tests. We also evaluate **STANCE** on simple collision scenarios captured with a handheld smartphone (Fig. 8), including tabletop rolling/sliding and two-object contact events. The model follows user-specified directions/speeds and preserves object identity across frames. In a *domino-like chain* example, it maintains consistent appearance for each piece (shape, texture, shading) while producing physically coherent interactions: contacts trigger sequential topples with plausible timing and momentum transfer, without pre-impact "hover" or bounce-back.

5 Limitations

Our approach is primarily limited to rigid body motions, struggling with complex non-rigid deformations. Additionally, regarding robustness, the model tolerates mild mask perturbations (e.g., small erosion or misalignment) well, but heavy noise in control arrows causes it to compromise between the noisy input and its learned priors, leading to weakened or rotated motion.

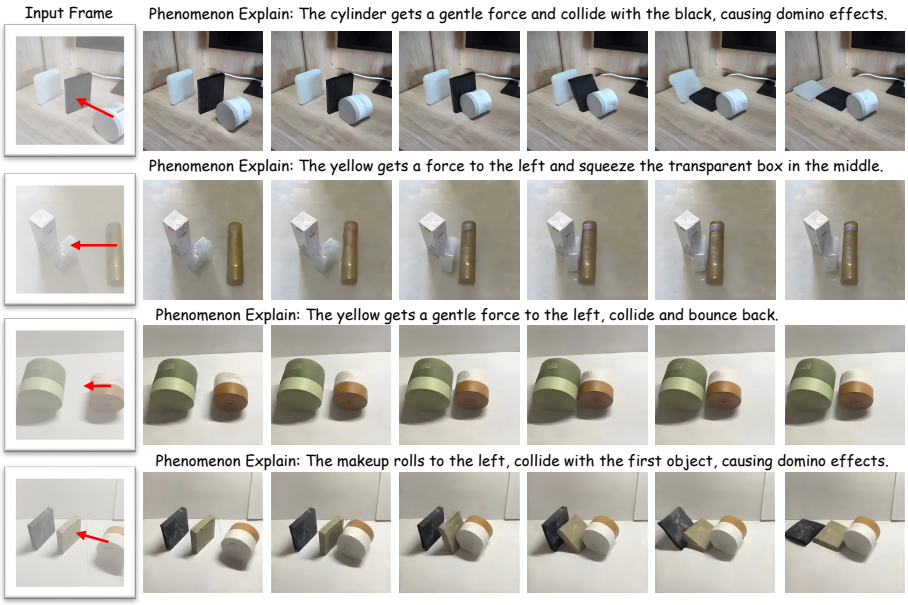

Fig. 8: Real-world demos. We evaluate **STANCE** on tabletop collisions shot with a hand-held phone. Given the first frame and user-specified initial velocities, STANCE follows the directions/speeds, preserves object identity, and produces physically coherent outcomes.

6 Conclusion

We present **STANCE**, a controllable image-to-video framework that turns sparse, human-editable hints into token-dense, pixel-aligned guidance for coherent motion synthesis. By introducing *Instance Cues* with camera-relative 2.5D depth and *Dense RoPE* for spatially addressable rotary embeddings, our method effectively resolves motion ambiguity and strengthens the control pathway. Coupled with a lightweight structural head that acts as a geometry witness, STANCE significantly improves temporal and interaction coherence while maintaining high visual quality.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
2. cerspense: zeroscope_v2. `https://huggingface.co/cerspense/zeroscope_v2_576w` (2023), accessed: 2023-02-03
3. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)

4. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. arXiv preprint arXiv:2401.09047 (2024)
5. Chen, T.S., Lin, C.H., Tseng, H.Y., Lin, T.Y., Yang, M.H.: Motion-conditioned diffusion model for controllable video synthesis. arXiv preprint arXiv:2304.14404 (2023)
6. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. Ieee (2022)
7. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3761 (2022)
8. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
9. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
10. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
11. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing (2025), https://arxiv.org/abs/2503.07598
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
13. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
14. Li, Y., Angel, M.C., Khan, S., Zhu, Y., Sun, J., Zhang, Y., Khan, F.S.: C-drag: Chain-of-thought driven motion controller for video generation. arXiv preprint arXiv:2502.19868 (2025)
15. Liu, Z., Yanev, A., Mahmood, A., Nikolov, I., Motamed, S., Zheng, W.S., Wang, X., Gool, L.V., Paudel, D.P.: Intragen: Trajectory-controlled video generation for object interactions (2024), https://arxiv.org/abs/2411.16804
16. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248 (2025)
17. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? arXiv preprint arXiv:2501.09038 (2025)
18. Namekata, K., Bahmani, S., Wu, Z., Kant, Y., Gilitschenski, I., Lindell, D.B.: Sg-i2v: Self-guided trajectory control in image-to-video generation. arXiv preprint arXiv:2411.04989 (2024)
19. Niu, M., Cun, X., Wang, X., Zhang, Y., Shan, Y., Zheng, Y.: Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In: European Conference on Computer Vision. pp. 111–128. Springer (2024)
20. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-

to-video generation with explicit motion modeling. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
21. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
22. Team, G.: Mochi 1. https://github.com/genmoai/models (2024)
23. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
24. Vainer, S., Boss, M., Parger, M., Kutsy, K., Nigris, D.D., Rowles, C., Perony, N., Donné, S.: Collaborative control for geometry-conditioned pbr image generation (2024), https://arxiv.org/abs/2402.05919
25. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
26. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)
27. Wu, W., Li, Z., Gu, Y., Zhao, R., He, Y., Zhang, D.J., Shou, M.Z., Li, Y., Gao, T., Zhang, D.: Draganything: Motion control for anything using entity representation. In: European Conference on Computer Vision. pp. 331–348. Springer (2024)
28. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., Jia, X.: Vlipp: Towards physically plausible video generation with vision and language informed physical prior (2025), https://arxiv.org/abs/2503.23368
29. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
30. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
31. Zhang, Z., Long, F., Qiu, Z., Pan, Y., Liu, W., Yao, T., Mei, T.: Motionpro: A precise motion controller for image-to-video generation (2025), https://arxiv.org/abs/2505.20287
32. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all (March 2024), https://github.com/hpcaitech/Open-Sora
33. Zhu, J.Y., Wu, J., Shi, Y., Zhou, T., Yang, D., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: DragAnything: Interactive point-based manipulation on the generative image manifold. arXiv preprint arXiv:2306.14435 (2023)