Event-Driven Video Generation

Chika Maduabuchi and Jindong Wang*

William & Mary, Williamsburg, VA, USA

Fig. 1: Representative text-conditioned video outputs produced by EVD. The samples highlight the intended effect of the event gate: latent updates are concentrated around prompt-relevant events, reducing pre-contact motion, missing interaction responses, and post-event drift.

Abstract. Current text-to-video models can make individual frames look convincing while still getting simple interactions wrong: objects move before contact, an intended action is skipped, a placed object keeps drifting, or a support relation breaks. Our starting point is that standard *frame-first* denoising updates every latent region at every step, even when the prompt implies that only a local interaction should be active. We introduce *Event-Driven Video Generation (EVD)*, a small DiT-compatible intervention that gives the sampler an explicit event signal. A lightweight head predicts token-level event activity; training losses tie that activity to latent state change; and event-gated sampling, with hysteresis and an early-step schedule, applies the update field mainly where an interaction is forming. On EVD-Bench, EVD improves human preference and VBench dynamics for *state persistence, spatial accuracy, support relations,* and *contact stability*, while keeping appearance quality comparable to the base model. The results suggest that a modest amount of event structure can correct several interaction failures that otherwise remain hidden behind good frame-level appearance. Project webpage: https://evd-project-website.pages.dev

Keywords: Text-to-video generation · Event grounding · Video diffusion transformers

* Corresponding author: jdw@wm.edu.

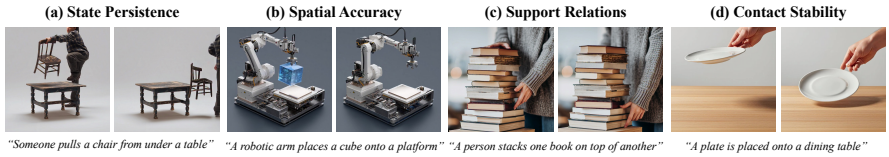

Fig. 2: Four interaction failures we observe in DiT-30B. The examples isolate mistakes that are easy to miss when judging only single-frame quality: (a) state persistence, where the chair keeps moving after the pull has ended; (b) spatial accuracy, where the cube never aligns with the target platform; (c) support relations, where the book appears stacked without a visible stacking action; and (d) contact stability, where the plate begins moving before hand–object contact.

1 Introduction

Text-to-video has improved quickly in frame realism, resolution, and clip length. Space–time diffusion models such as Lumiere [1] can synthesize coherent clips in one pass, and large diffusion-transformer systems such as Movie Gen [27] and Step-Video-T2V [20] scale latent representations and training recipes to longer, higher-fidelity videos. Open and semi-open systems, including Hunyuan-Video [14], Open-Sora [39], and Wan [32], together with faster samplers such as Pyramidal Flow Matching [12] and T2V-Turbo/T2V-Turbo-v2 [15, 16], make these capabilities more broadly available. As the frames become more convincing, the remaining errors are often no longer about texture or sharpness; they are about whether an interaction actually happened in the right place and in the right order.

2 Motivation

These gains do not remove a basic interaction problem. A clip can be smooth locally while the cause–effect story is wrong: effects appear before their causes, contact is missing, or the object keeps changing after the action should be over. Recent evaluation suites separate appearance from motion, physics, and temporal consistency, making the same point quantitatively: good-looking frames do not guarantee realistic dynamics (VBench [10], VBench++ [11], VBench-2.0 [37]). In our experiments, this is also where large models remain fragile: they may score well on appearance and still fail on interaction grounding, which suggests that scaling alone is not a reliable fix [28].

Recent systems improve dynamics with stronger architectures and training/inference recipes (CogVideoX [34], Open-Sora STDiT [39], Hunyuan-Video [14], Step-Video-T2V [20]), and with motion-aware objectives or inference-time steering that reduce the appearance–motion imbalance (VideoJAM [4]). For the interaction failures in Fig. 2, however, a common issue remains: the sampler is still mostly "frame-first," updating the latent state everywhere at every step. That makes it easy for a locally plausible update to

occur before the relevant contact, to miss the actual state change, or to keep drifting after the event. We group these errors into State Persistence, Spatial Accuracy, Support Relations, and Contact Stability. The mechanism we want is correspondingly simple: the model should know whether an interaction is active *here* and *now*, and it should use that signal to decide where the latent is allowed to change.

3 Related Work

Video generation and evaluation. Recent text-to-video systems have advanced through space–time diffusion, large video DiT backbones, temporal autoencoders, and efficient flow/diffusion sampling [1, 12, 14, 20, 27, 32, 34, 39]. The remaining interaction errors are often specific: motion starts too early, contact is weak, support relations are skipped, or the post-event state keeps moving. Evaluation suites such as VBench, VBench++, VBench-2.0, WorldSimBench, T2V-CompBench, and NeuS-V make a similar distinction between visual quality and temporally faithful or physically meaningful generation [10, 11, 28, 30, 31, 37]. EVD addresses these errors by coupling latent updates to prompt-relevant event activity.

Event structure, motion steering, and editing. The closest event-centric prior work is GEST, which represents visual/language stories as Graphs of Events in Space and Time [24], and GEST-Engine, which executes formal GEST specifications to synthesize controllable multi-actor videos with dense spatiotemporal annotations [5]. These works motivate events as an abstraction, but rely on symbolic/event-graph specifications or simulation-style control. EVD instead learns token-aligned event activity inside a pretrained video DiT and uses it to gate the sampled direction field. EVD also differs from motion-steering and editing/correction pipelines such as VideoJAM [4], StreamDiffusion [13], StreamV2V [17], and ObjectAlign [25]: it is a text-to-video generation-time mechanism that keeps the solver, decoder, NFE, and output-selection protocol unchanged while modifying only event-grounded training and the direction field.

4 Contributions

We propose *Event-Driven Video Generation (EVD)* as a small add-on for pretrained video DiTs rather than a replacement backbone. EVD makes three changes. First, it attaches a lightweight event head to DiT token features and predicts a token-aligned activity map. Second, it trains the model so latent updates are tied to that activity: inactive regions are discouraged from changing, while active interaction regions receive more stable updates. Third, at sampling time, it gates the solver direction field with a hysteresis-and-schedule rule, so early event formation is allowed but late spurious drift is damped. This keeps the solver family and decoder unchanged, which makes the method compatible

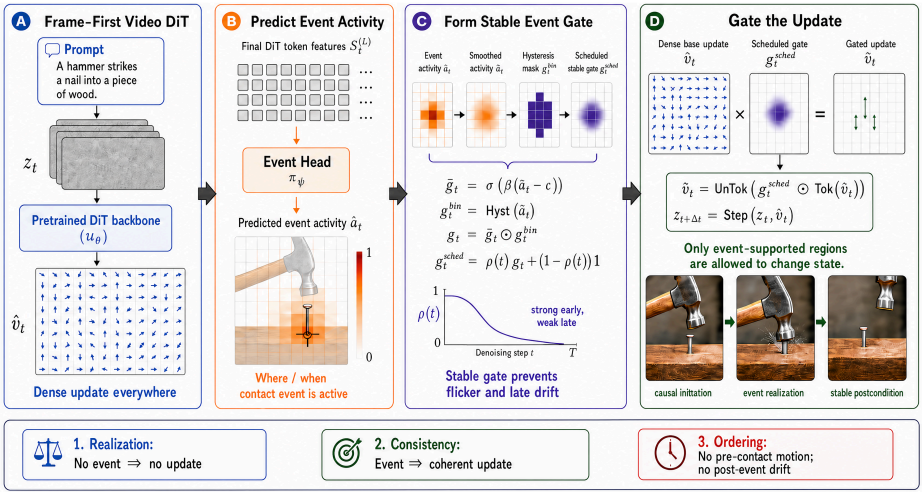

Fig. 3: Overview of Event-Driven Video Generation (EVD). Given a noised latent video z_t and prompt y, the DiT predicts its usual update field $\widehat{v}_t$. A lightweight event head predicts token-aligned activity $\hat{a}_t$, which we smooth and convert into a gate using soft activation, hysteresis, and an early-step schedule. The solver and decoder are unchanged; the intervention is that the update field is no longer applied uniformly everywhere. During training, the event losses tie active regions to state change and discourage unrelated regions from moving.

with modern DiT video systems [20, 27, 34, 39]. The main limitation we observe is also specific: if the event signal is weak at the operating resolution, as in small contacts, occlusion, or cluttered multi-object scenes, EVD can under-localize the interaction. This points to object-centric or contact-aware event cues as natural next steps.

5 EVD

This section describes how EVD changes a pretrained video DiT without replacing its backbone. The DiT still predicts a latent update field, but we add a token-level event pathway that decides where that update should be trusted. We start with the latent-video interface (Sec. 5.1), define the event head and gated update rule (Sec. 5.2), then specify the soft/hysteretic gate and the resulting direction field (Sec. 5.2). The training objective adds event realization, consistency, ordering, and time-weighting terms on top of Flow Matching (Sec. 5.3), and inference uses the same idea under CFG with a scheduled gate (Sec. 5.4). The accompanying Supplementary Material gives the exact training and sampling pseudocode, while Sec. 5.5 lists the hyperparameters used in the experiments.

5.1 Method: Event-Driven Video Generation (EVD)

Figure 3 summarizes EVD, from token-aligned event activity prediction and stable gate formation to gated latent updates during sampling.

Latent video representation and notation We consider text-conditioned video generation. A video clip is denoted by $x \in \mathbb{R}^{T\times H\times W\times 3}$ with T frames. Following standard practice in large-scale video diffusion/flow models, we generate in a compressed latent space using a temporal video autoencoder. Let $E(\cdot)$ and $D(\cdot)$ denote the encoder and decoder, and define the latent video

$$z = E(x), \qquad x \approx D(z), \tag{1}$$

where $z \in \mathbb{R}^{T'\times H'\times W'\times C}$ has reduced spatiotemporal resolution.

Let y denote the text prompt, encoded by a frozen text encoder into a sequence of embeddings. We denote the DiT backbone by $u_\theta(\cdot)$, parameterized by θ, which operates on noised latents and conditions on (y, t), where $t \in [0, 1]$ is the continuous diffusion/flow time.

Noise model and training target. We use a continuous-time Flow Matching formulation in latent space [18, 21, 23]. Given a clean latent z_1 and noise $z_0 \sim \mathcal{N}(0, I)$, we sample $t \sim \mathcal{U}[0, 1]$ and form the interpolated latent

$$z_t = t\, z_1 + (1 - t)\, z_0, \tag{2}$$

with velocity target

$$v_t = \frac{dz_t}{dt} = z_1 - z_0. \tag{3}$$

The backbone predicts $\widehat{v}_t = u_\theta(z_t, y, t)$. We keep this interface fixed and modify how the model represents and applies *event-driven* state changes in the subsequent subsections.

DiT backbone and conditioning interface EVD is built on a pretrained video Diffusion Transformer (DiT) operating in latent space. Given $z_t \in \mathbb{R}^{T'\times H'\times W'\times C}$, we patchify z_t into spatiotemporal tokens and map them to the model width d, yielding a token sequence $\mathbf{s}_t \in \mathbb{R}^{N\times d}$, where N is the number of spatiotemporal patches. The DiT applies L transformer blocks with spatiotemporal self-attention and text cross-attention, producing final features $\mathbf{s}_t^{(L)}$.

The DiT backbone outputs a prediction of the Flow Matching velocity (Eq. (3)):

$$\widehat{v}_t \;=\; u_\theta(z_t, y, t). \tag{4}$$

We do not modify the backbone architecture or its tokenization; EVD introduces an additional lightweight *event pathway* that predicts event activity aligned with the same token grid and uses it to gate updates during training and sampling (Secs. 5.2–5.4).

Sampling interface. Sampling evolves a latent trajectory $\{z_{t_k}\}_{k=0}^{K}$ along a monotone time grid $0 = t_0 < \cdots < t_K = 1$. At each step, the sampler queries the backbone to obtain a direction field and updates z_{t_k} using a solver step (Euler/Heun/DPM-style [38]). EVD is compatible with any such solver because it only changes the direction field passed to the solver.

Classifier-free guidance (CFG) We use classifier-free guidance (CFG) to improve prompt adherence. At each sampling step t_k, we evaluate the backbone with the prompt y and with a null prompt $\varnothing$, and form the guided direction field

$$\widehat{v}^{\mathrm{cfg}}(z_{t_k}, y, t_k) \;=\; (1 + w_{\mathrm{cfg}})\, u_\theta(z_{t_k}, y, t_k) \;-\; w_{\mathrm{cfg}}\, u_\theta(z_{t_k}, \varnothing, t_k), \tag{5}$$

where $w_{\mathrm{cfg}} \geq 0$ is the guidance scale. EVD applies event gating *after* forming $\widehat{v}^{\mathrm{cfg}}$, so prompt adherence is preserved while spurious, event-inconsistent dynamics are suppressed (Sec. 5.4).

Why frame-first generation fails on interactions Although modern video generators can produce locally smooth motion [1, 4, 9, 14, 20, 27, 33, 39], they often violate basic causal structure in simple interactions: effects appear without causes (e.g., objects move before contact), causes occur without coherent effects (e.g., an interaction is implied but the state does not respond), and post-interaction states drift instead of settling. These errors are particularly salient in prompts involving contact, support, constrained mechanisms, or material transfer, and they map directly to the four failure categories used in our analysis: *State Persistence*, *Spatial Accuracy*, *Support Relations*, and *Contact Stability*.

5.2 Event-Driven Video Generation

Core idea: event-gated state updates EVD models a video as persistent latent state punctuated by discrete interaction events. At diffusion/flow time t, we introduce an *event representation* e_t aligned with the DiT token grid, and derive from it an event gate $g_\psi(e_t, t) \in [0, 1]^N$ (token-wise, broadcast across channels). The gate modulates the backbone direction field so that state updates occur only when justified by an active interaction:

$$\Delta z_t \;=\; \mathrm{UnTok}\Big(g_\psi(e_t, t) \odot \mathrm{Tok}\big(u_\theta(z_t, y, t)\big)\Big). \tag{6}$$

When the event signal indicates "no interaction" (gate near zero), the update is suppressed and the state remains stable; when an event is active (gate near one), the update proceeds normally. This single mechanism targets both common degeneracies: *missing events* (state changes without a visible interaction) and *ghost events* (an implied interaction without a coherent state change).

In the remainder of this section, we specify (i) how e_t and g_ψ are instantiated, (ii) the event-grounded training objective, and (iii) the event-driven sampling procedure with CFG.

Event representation and event head EVD uses a token-aligned event field to localize interactions in space and time. Let $\mathbf{s}_t^{(L)} \in \mathbb{R}^{N\times d}$ be the final DiT token features at time t. We attach a lightweight event head π_ψ that predicts an event field

$$\hat{e}_t \;=\; \pi_\psi(\mathbf{s}_t^{(L)}, t) \;\in\; \mathbb{R}^{N\times C_e}, \tag{7}$$

with a small channel budget C_e (we use $C_e = 1$ in the main method). The first channel is interpreted as an *event activity* logit, and we obtain a token-wise activity probability via

$$\hat{a}_t \;=\; \sigma(\hat{e}_t^{(1)}) \;\in\; [0,1]^N, \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid and $\hat{e}_t^{(1)}$ denotes the activity channel.

Zero-impact initialization. To preserve pretrained DiT behavior at the start of fine-tuning, we initialize π_ψ to near-zero output so that $\hat{a}_t \approx 0$ initially, and the model reduces to the base backbone before learning event structure.

Spatial smoothing. Event activity can be spatially fragmented due to noise. We apply a lightweight smoothing operator $\mathcal{S}$ over the spatial patch grid (per frame) and use $\tilde{a}_t = \mathcal{S}(\hat{a}_t)$ in all gating computations. In our main setting, $\mathcal{S}$ is a 3×3 average filter.

Event gate: soft activation with hysteresis We convert the smoothed activity $\tilde{a}_t \in [0,1]^N$ into a stable, token-wise gate $g_t \in [0,1]^N$ that controls whether each token is allowed to update. EVD uses *soft activation* to avoid brittle thresholding and *hysteresis* to prevent flickering event boundaries.

Soft activation. We first form a soft gate centered between the on/off thresholds:

$$\bar{g}_t \;=\; \sigma\Big(\beta\big(\tilde{a}_t - \tfrac{\tau_{\text{on}}+\tau_{\text{off}}}{2}\big)\Big), \tag{9}$$

where $\beta > 0$ controls sharpness and $\tau_{\text{on}} > \tau_{\text{off}}$ define the hysteresis band.

Hysteresis update. We maintain a binary state gate $g_t^{\text{bin}} \in \{0,1\}^N$ with token-wise update:

$$g_{t,i}^{\text{bin}} \;=\; \begin{cases} 1, & \tilde{a}_{t,i} \geq \tau_{\text{on}}, \\ 0, & \tilde{a}_{t,i} \leq \tau_{\text{off}}, \\ g_{t^-,i}^{\text{bin}}, & \text{otherwise}, \end{cases} \tag{10}$$

where t^- denotes the previous sampling step (or previous iteration in the discretized schedule) and i indexes tokens. Finally, we combine soft and hysteresis gates to obtain the effective gate used for modulation:

$$g_t \;=\; \bar{g}_t \odot g_t^{\text{bin}}, \tag{11}$$

Here g_t^{bin} provides stable on/off event activation (prevents flicker), while $\bar{g}_t$ smoothly scales update magnitude within active regions; scheduled gating is applied afterward during sampling (Sec. 5.4).

In practice this reduces to using the hysteresis state for stability while retaining smooth transitions through $\bar{g}_t$. (The exact implementation used in our experiments is provided as pseudocode in the Supplementary Material.)

Event-gated update field Given the backbone prediction $\widehat{v}_t = u_\theta(z_t, y, t)$ and the event gate $g_t \in [0,1]^N$, EVD forms an event-gated direction field by modulating the patchified backbone output and unpatchifying back to latent space:

$$\widetilde{v}_t \;=\; \mathrm{UnTok}\Big(g_t \odot \mathrm{Tok}\big(\widehat{v}_t\big)\Big). \tag{12}$$

We then pass $\widetilde{v}_t$ to the same solver used by the base model. Because EVD only changes the direction field, it is compatible with any ODE sampler (Euler/Heun/DPM-style) used for DiT video generation.

5.3 Training Objective

We keep the base Flow Matching objective (Eqs. (2)–(3)) and add event-grounded terms that couple event activity to state evolution. Let $\widehat{v}_t = u_\theta(z_t, y, t)$ be the predicted velocity and $\Delta_t = \mathrm{Tok}(\widehat{v}_t)$ its token form.

Base Flow Matching loss The base loss matches the predicted velocity to the target $v_t = z_1 - z_0$:

$$\mathcal{L}_{\text{base}} \;=\; \mathbb{E}_{z_1, z_0, t, y}\Big[\big\|u_\theta(z_t, y, t) - (z_1 - z_0)\big\|_2^2\Big]. \tag{13}$$

Event realization loss To prevent *missing events* (state changes without an active interaction), we penalize update energy in tokens where the event activity is low:

$$\mathcal{L}_{\text{real}} \;=\; \mathbb{E}\Big[\big\|(1 - \tilde{a}_t) \odot \Delta_t\big\|_2^2\Big], \tag{14}$$

where $\tilde{a}_t$ is the smoothed event activity (Sec. 5.2). This term forces the model to either (i) predict an active event where a state change is required, or (ii) suppress the state change when no event is present.

Event consistency loss To prevent *ghost events* and reduce jitter during interactions, we enforce that state updates under active events are locally consistent across nearby diffusion/flow times. For each training sample we draw a second time $t' = \mathrm{clip}(t + \delta, 0, 1)$ with $\delta \sim \mathcal{U}[-\Delta, \Delta]$, construct $z_{t'} = t' z_1 + (1 - t') z_0$, and compute $\Delta_t = \mathrm{Tok}(u_\theta(z_t, y, t))$, $\Delta_{t'} = \mathrm{Tok}(u_\theta(z_{t'}, y, t'))$, with corresponding activities $\tilde{a}_t$ and $\tilde{a}_{t'}$. We then minimize the event-masked discrepancy:

$$\mathcal{L}_{\text{cons}} \;=\; \mathbb{E}\Big[\big\|\tilde{a}_t \odot \Delta_t - \tilde{a}_{t'} \odot \Delta_{t'}\big\|_2^2\Big]. \tag{15}$$

Intuitively, once an interaction is active, the model should not oscillate between incompatible update directions across infinitesimally close times; $\mathcal{L}_{\text{cons}}$ encourages stable, directed state evolution during the event.

Ordering and termination loss EVD additionally enforces a simple causal ordering: motion should not occur before event initiation and should decay after termination. Using the same thresholds that define the hysteresis band ($\tau_{\text{on}} > \tau_{\text{off}}$), we suppress update energy in low-activity regions:

$$\mathcal{L}_{\text{order}} \;=\; \mathbb{E}\Big[\big\|\mathbf{1}[\tilde{a}_t < \tau_{\text{on}}] \odot \Delta_t\big\|_2^2 + \big\|\mathbf{1}[\tilde{a}_t < \tau_{\text{off}}] \odot \Delta_t\big\|_2^2\Big], \tag{16}$$

where $\mathbf{1}[\cdot]$ is applied token-wise. The first term discourages pre-event motion (improving *Contact Stability*); the second term suppresses residual updates after the model indicates the event is off (improving *State Persistence*).

Time-weighted objective Event grounding matters most at early diffusion/flow times that determine coarse dynamics. We therefore apply a time weight

$$w(t) = \mathbf{1}[t \le t^\star_{\text{loss}}] + \exp\big(-\kappa(t - t^\star_{\text{loss}})\big)\mathbf{1}[t > t^\star_{\text{loss}}], \tag{17}$$

and optimize the total objective

$$\mathcal{L} \;=\; \mathcal{L}_{\text{base}} + w(t)\Big(\lambda_{\text{real}}\mathcal{L}_{\text{real}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{order}}\mathcal{L}_{\text{order}}\Big). \tag{18}$$

The complete training procedure is provided in the Supplementary Material.

5.4 Inference: Event-Driven Sampling

At inference, EVD uses the same sampler and decoder as the base DiT model, but replaces the direction field passed to the solver with an event-gated field. This change directly suppresses pre-contact motion and post-interaction drift while preserving prompt adherence via CFG.

Scheduled event gating Event grounding is most important early in sampling, where coarse motion and interaction structure are established. We therefore apply event gating strongly for early steps and anneal it later. Given sampling time t_k, we define

$$\rho(t_k) = \begin{cases} 1, & t_k \le t^\star, \\ 1 - \dfrac{t_k - t^\star}{1 - t^\star}, & t_k > t^\star, \end{cases} \tag{19}$$

and combine it with the gate g_k to obtain the scheduled gate

$$g_k^{\text{sched}} \;=\; \rho(t_k)\, g_k + (1 - \rho(t_k))\, \mathbf{1}, \tag{20}$$

where $\mathbf{1}$ is the all-ones gate (no gating). When $\rho = 1$, EVD applies full event gating; when $\rho = 0$, sampling reduces to the base model.

Event-driven solver update At each step we compute the hysteresis gate g_k^{bin} (Eq. (10)) and the soft gate $\bar{g}_k$ (Eq. (9)), set $g_k = \bar{g}_k \odot g_k^{\text{bin}}$ (Eq. (11)), and then apply the scheduled gate g_k^{sched} (Eq. (20)). We then pass the gated direction field to the base solver:

$$\widetilde{v}_{t_k} \;=\; \mathrm{UnTok}\Big(g_k^{\text{sched}} \odot \mathrm{Tok}\big(\widehat{v}^{\text{cfg}}(z_{t_k}, y, t_k)\big)\Big), \tag{21}$$

$$z_{t_{k+1}} \;=\; \mathrm{Step}(z_{t_k}, \widetilde{v}_{t_k}, t_k, t_{k+1}), \tag{22}$$

where $\mathrm{Step}(\cdot)$ is any solver step used by the base model (Euler/Heun/DPM-style). The complete sampling procedure is provided in the Supplementary Material.

5.5 Practical settings

For all experiments, we use the same latent clip format and sampling interface described in Sec. 5.1. Unless otherwise stated, we use $K = 50$ sampling steps with CFG scale $w_{\text{cfg}} = 4.0$. Event gating uses $\beta = 12.0$ and hysteresis thresholds $\tau_{\text{on}} = 0.62$, $\tau_{\text{off}} = 0.38$, with 3×3 spatial smoothing on the patch grid. We apply scheduled gating with cutoff $t^\star = 0.60$, i.e., full gating for early steps and linear annealing thereafter (Sec. 5.4). For training, we set $t^\star_{\text{loss}} = 0.60$ and $\kappa = 6$ in the time-weight $w(t)$ (Eq. (17)), use event dropout $p_e = 0.25$, and optimize Eq. (18) with $\lambda_{\text{real}} = 0.12$, $\lambda_{\text{cons}} = 0.08$, and $\lambda_{\text{order}} = 0.03$. These values are sufficient to reproduce the qualitative behaviors in our figures and the quantitative gains on EVD-Bench; the Supplementary Material provides compute/data details, extended ablations, and sensitivity analyses.

6 Experiments

We evaluate EVD on prompts where the important question is not just whether the frames look plausible, but whether the interaction happens in the right order. The section covers the EVD-Bench setup, qualitative comparisons, human and automatic metrics, and the ablations needed to separate event grounding from a generic motion mask.

6.1 Setup

We evaluate on EVD-Bench, a curated set of 150 short interaction-centric prompts that stress causal event realization, grouped into four failure categories used throughout (Fig. 2): *State Persistence*, *Spatial Accuracy*, *Support Relations*, and *Contact Stability*. Our primary baselines are pretrained DiT-4B and DiT-30B, with DiT-4B+EVD and DiT-30B+EVD applying EVD as an additive modification (lightweight event head + event-driven training/sampling) without changing transformer blocks; Fig. 6 additionally includes strong external video generators for qualitative reference. All methods generate 128-frame

clips at 24 fps with a 256×256 base generation resolution in the temporal-autoencoder latent/decoder space (upsampled to 720p for visualization), using a matched solver, step budget K (NFE), and CFG scale w_{cfg}; critically, EVD changes only the direction field passed to the sampler (event gating) and uses identical decoding with no post-hoc filtering. We report automatic VBench *Appearance*/*Dynamics* and human 2AFC preferences over *Text Faithfulness*, *Quality*, and *Dynamics*, running each model once per prompt with a fixed seed and evaluating the first sample (no cherry-picking). Human-evaluation details, EVD-Bench construction and leakage safeguards, and closed-source normalization are provided in the Supplementary Material.

6.2 Qualitative Results

Fig. 4 shows six EVD samples: ball-through-hoop, sliding door, sponge press/release, trash-can lid open/close, two-person pass, and liquid pouring.[1] We use these examples to check the concrete behavior of the method. The motion starts after the triggering action, contact regions stay spatially plausible, and the scene usually settles instead of continuing to drift after the event.

Fig. 4: Representative EVD samples. The six prompts cover target-directed motion, a constrained mechanism, deformation and recovery, gravity-mediated closure, a two-person handoff, and liquid transfer. The examples are meant to show the local effect of the event gate: motion is concentrated near the active interaction and reduced once the intended state change has happened.

[1] These are exactly the six prompts shown in Fig. 4.

Table 1: Comparison of EVD with video generation baselines on EVD-Bench. Human evaluation reports the percentage of pairwise votes favoring EVD; automatic metrics are computed using VBench. *TF* denotes Text Faithfulness, *Qual.* denotes Overall Quality, *Dyn.* denotes Dynamics, and *App.* denotes Appearance; higher is better for all columns.

(a) Prior video generation baselines

	Human Eval			Auto. Metrics	
Method	**TF**	**Qual.**	**Dyn.**	**App.**	**Dyn.**
CogVideo2B	80.2	88.1	89.7	69.8	87.6
CogVideo5B	65.4	73.8	72.5	72.3	89.2
PyramidFlow	75.8	82.4	81.1	73.9	88.5
DiT-4B	70.6	76.8	80.3	75.4	78.9
DiT-4B+EVD	**88.9**	**91.3**	**96.4**	**76.2**	**94.8**

(b) Large-scale video generators

	Human Eval			Auto. Metrics	
Method	**TF**	**Qual.**	**Dyn.**	**App.**	**Dyn.**
Kling 3.0 Pro	61.8	67.4	72.9	78.6	92.6
Runway Gen-4.5	64.7	71.2	76.8	76.9	91.4
Veo 3.1	66.9	73.5	78.4	77.2	91.8
Sora 2 Pro	63.5	69.8	74.1	77.8	90.9
Mochi 1	58.2	63.7	70.3	72.6	88.8
DiT-30B	72.4	76.1	79.5	73.8	88.7
DiT-30B+EVD	**89.7**	**92.4**	**97.1**	**78.1**	**95.7**

6.3 Quantitative Results on EVD-Bench

Table 1 combines the human 2AFC study with VBench scores on EVD-Bench. Human raters prefer DiT+EVD over the corresponding DiT baseline for *Text Faithfulness*, *Overall Quality*, and most strongly *Dynamics*. The automatic metrics show the same pattern we want to see: VBench *Dynamics* increases, while *Appearance* stays close to the base model. This matters because many failures in Fig. 2 are not low-level visual artifacts; they are incorrect or unstable state updates inside otherwise realistic-looking clips.

Stress tests and diagnostics. Beyond EVD-Bench, we evaluate fixed-seed compositional/temporal and simultaneous-event subsets drawn from T2V-CompBench [31] and NeuS-V [30], compare against recent open-source generators (Wan/Hunyuan) [14, 32], and report confidence intervals and motion-mask controls in the Supplementary Material. Fig. 5 further validates that EVD activity localizes to prompt-relevant interaction regions rather than diffuse background motion: the pseudo-target, learned activity, and final gate concentrate around placement or contact and material-transfer events while suppressing inactive regions. This supports the central claim that EVD learns an event-grounded update signal, not merely a generic motion mask.

6.4 Baseline Comparisons

Fig. 6 compares four interaction prompts. The baselines often look realistic in isolated frames, but the event itself is misplaced: an object begins changing before contact, the deformation or constraint response is too weak, or the scene keeps drifting after the interaction should have ended. We see this for material transfer (coffee filling), compliance (pillow compression), constraint enforcement (rope straightening), and multi-agent transitions (elevator door opening and people stepping inside). EVD does not solve every detail, but in these examples the main state change is better aligned with the trigger and has a more stable postcondition.

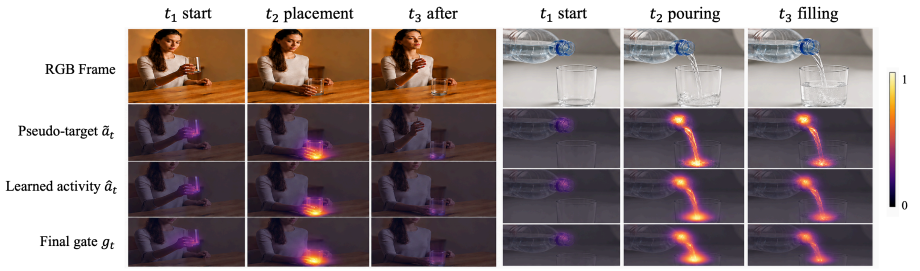

Fig. 5: Pseudo-target and gate localization diagnostics. We visualize placement/contact (left) and pouring/material transfer (right) at three time points. The first row shows RGB frames; lower rows overlay event signals on the corresponding frames. The pseudo-target $\tilde{a}_t$ is the self-supervised event target extracted from localized change, the learned activity $\hat{a}_t$ is the event-head prediction, and the final gate g_t is the smoothed, hysteretic, scheduled gate used to modulate latent updates. The three signals are aligned but not identical because they correspond to different stages of the EVD pipeline: supervision, prediction, and update gating. Activity concentrates at the glass–table contact and along the bottle-mouth, stream, and receiving-glass transfer path, while inactive/background regions remain suppressed.

Fig. 6: Qualitative comparison with video generation baselines. We compare EVD against Movie Gen, Sora, and DiT-30B on representative prompts involving soft-body deformation, flexible-object dynamics, structured scene interactions, and liquid transfer. The comparison highlights where the event gate changes the result: the baseline clips often miss the contact response or keep changing state after the intended interaction, while EVD better localizes the state change around the active event.

Table 2: Efficiency and overhead. EVD adds a lightweight event head and gating logic while keeping the same sampler, steps (K=50), and CFG structure (2 DiT evaluations per step).

Model	Total Params	Added Params	Added (%)	Train Throughput	Inference Overhead	DiT evals/step	Steps (K)
DiT-4B	4.0B	–	–	1.00×	1.00×	2 (CFG)	50
DiT-4B+EVD	4.0B	6.5M	0.16%	0.98×	1.02×	2 (CFG)	50
DiT-30B	30.0B	–	–	1.00×	1.00×	2 (CFG)	50
DiT-30B+EVD	30.0B	12.0M	0.04%	0.97×	1.02×	2 (CFG)	50

Table 3: EVD ablations on EVD-Bench (DiT-4B backbone). Human Eval reports the *percentage of 2AFC votes favoring the full DiT-4B+EVD model* over each ablated variant (higher is better for EVD). **Auto. Metrics** are computed using VBench. All variants use identical sampling settings (solver, NFE, CFG scale) and the same prompt set.

	Settings					Human Eval (EVD wins %)			Auto. Metrics	
Variant	Real	Cons	Order	Gate	Sched.	Text Faith.	Quality	**Dynamics**	Appearance	**Dynamics**
DiT-4B (no EVD)	No	No	No	No	–	70.6	76.8	80.3	75.4	78.9
w/o event realization	No	Yes	Yes	Yes	Anneal	65.2	69.4	78.8	75.9	91.2
w/o event consistency	Yes	No	Yes	Yes	Anneal	61.3	65.0	74.2	76.0	92.1
Training-only (no gating)	Yes	Yes	Yes	No	Off	63.0	66.8	77.1	76.1	90.5
Inference-only (no event losses)	No	No	No	Ext.	Anneal	70.5	74.9	86.0	75.6	84.0
Disable gating & event use at inference	Yes	Yes	Yes	No	–	67.8	71.6	82.4	75.5	86.7
No schedule (const. gate)	Yes	Yes	Yes	Yes	Const. (1.0)	55.7	58.9	60.5	75.7	94.1
No schedule (weak const. gate)	Yes	Yes	Yes	Yes	Const. (0.5)	57.9	61.0	65.8	76.1	93.6
DiT-4B + EVD (full)	Yes	Yes	Yes	Yes	Anneal	**88.9**	**91.3**	**96.4**	**76.2**	**94.8**

6.5 Ablations and Sensitivity

Component ablations. Table 3 breaks down the EVD components on EVD-Bench. Removing *realization* mostly brings back non-causal initiation (*Contact Stability*); removing *consistency* makes interactions noisier and less stable (*State Persistence*); and removing *ordering* weakens the settling behavior after an event. Training-only EVD and inference-only EVD recover only part of the gain. The controls show that motion masking alone is not enough; the useful part is training the model to align its own event signal with the latent update. We also find that constant gating without annealing hurts preference, which is why the full model uses the scheduled gate. Details of the motion-mask audit are provided in the Supplementary Material.

Efficiency and overhead. EVD keeps the backbone, solver, decoder, and sampling budget (K/NFE) fixed; the extra work is only the event head and direction-field gate. Table 2 shows that this adds negligible parameters and about 1.02× inference overhead.

Hyperparameter sensitivity. EVD is broadly robust to K (NFE), CFG scale w_{cfg}, and gating hyperparameters ($\beta, \tau_{\text{on}}, \tau_{\text{off}}, t^{\star}$); the full sensitivity sweep is provided in the Supplementary Material.

7 Conclusion

We introduced Event-Driven Video Generation (EVD), a small modification to pretrained video DiTs that ties latent updates to predicted event activity. The main empirical takeaway is that the model does not need a new solver or decoder to reduce several interaction errors: when the event head is trained together with the latent update and used during sampling, EVD improves human preference and automatic dynamics metrics on EVD-Bench while leaving appearance largely unchanged.

Limitations. EVD is least reliable when the event is hard to localize at the model's operating resolution: small or occluded contacts, cluttered multi-object interactions, thin fluid effects, or dominant camera motion. In those cases, both the pseudo-targets and the learned event head can become ambiguous, and the gate may either suppress useful motion or allow unrelated motion through. Future work should make the event signal more object- and contact-aware, and may also benefit from stronger motion disentanglement or higher-resolution latent/video backbones.

Acknowledgements

The authors thank NSF and NCSA for computational support. This paper used generative AI tools for language proof.

References

1. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., Li, Y., Rubinstein, M., Michaeli, T., Wang, O., Sun, D., Dekel, T., Mosseri, I.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. SA '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3680528.3687614, https://doi.org/10.1145/3680528.3687614
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D.: Stable video diffusion: Scaling latent video diffusion models to large datasets. ArXiv **abs/2311.15127** (2023), https://api.semanticscholar.org/CorpusID:265312551
3. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to Follow Image Editing Instructions . In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18392–18402. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.01764, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01764
4. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings

of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 7595–7616. PMLR (13–19 Jul 2025), https://proceedings.mlr.press/v267/chefer25a.html
5. Cudlenco, N., Masala, M., Leordeanu, M.: [tiny paper] GEST-engine: Controllable multi-actor video synthesis with perfect spatiotemporal annotations. In: ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling (2026), https://openreview.net/forum?id=uUofPYVMZH
6. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI
7. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2790–2799. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/houlsby19a.html
8. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
9. Huang, H., Ma, G., Duan, N., Chen, X., Wan, C., Ming, R., Wang, T., Wang, B., Lu, Z., Li, A., Zeng, X., Zhang, X., Yu, G., Yin, Y., Wu, Q., Sun, W., An, K., Han, X., Sun, D., Ji, W., Huang, B., Li, B., Wu, C., Huang, G., Xiong, H., He, J., Wu, J., Yuan, J., Wu, J., Liu, J., Guo, J., Tan, K., Chen, L., Chen, Q., Sun, R., Yuan, S., Yin, S., Liu, S., Chen, W., Dai, Y., Luo, Y., Ge, Z., Guan, Z., Song, X., Zhou, Y., Jiao, B., Chen, J., Li, J., Zhou, S., Zhang, X., Xiu, Y., Zhu, Y., Shum, H.Y., Jiang, D.: Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model (2025), https://arxiv.org/abs/2503.11251
10. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21807–21818 (June 2024)
11. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., Wang, Y., Chen, X., Chen, Y.C., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models . IEEE Transactions on Pattern Analysis & Machine Intelligence **48**(03), 3268–3285 (Mar 2026). https://doi.org/10.1109/TPAMI.2025.3633890, https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3633890
12. Jin, Y., Sun, Z., Li, N., Xu, K., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., MU, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=66NzcRQuOq
13. Kodaira, A., Xu, C., Hazama, T., Yoshimoto, T., Ohno, K., Mitsuhori, S., Sugano, S., Cho, H., Liu, Z., Tomizuka, M., Keutzer, K.: Streamdiffusion: A pipeline-level solution for real-time interactive generation. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12371–12380 (2025). https://doi.org/10.1109/ICCV51701.2025.01150
14. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C.,

Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models (2025), https://arxiv.org/abs/2412.03603

15. Li, J., Feng, W., Fu, T.J., Wang, X., Basu, S., Chen, W., Wang, W.Y.: T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=53daI9kbvf
16. Li, J., Long, Q., Zheng, J., Gao, X., Piramuthu, R., Chen, W., Wang, W.Y.: T2v-turbo-v2: Enhancing video model post-training through data, reward, and conditional guidance design. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=BZwXMqu4zG
17. Liang, F., Kodaira, A., Xu, C., Tomizuka, M., Keutzer, K., Marculescu, D.: Looking backward: Streaming video-to-video translation with feature banks. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 46425–46445 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/7280f65ed571b7b28321f2c7cf4c60c8-Paper-Conference.pdf
18. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
19. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 423–439. Springer Nature Switzerland, Cham (2022)
20. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., Zhou, Y., Sun, D., Zhou, D., Zhou, J., Tan, K., An, K., Chen, M., Ji, W., Wu, Q., Sun, W., Han, X., Wei, Y., Ge, Z., Li, A., Wang, B., Huang, B., Wang, B., Li, B., Miao, C., Xu, C., Wu, C., Yu, C., Shi, D., Hu, D., Liu, E., Yu, G., Yang, G., Huang, G., Yan, G., Feng, H., Nie, H., Jia, H., Hu, H., Chen, H., Yan, H., Wang, H., Guo, H., Xiong, H., Xiong, H., Gong, J., Wu, J., Wu, J., Wu, J., Yang, J., Liu, J., Li, J., Zhang, J., Guo, J., Lin, J., Li, K., Liu, L., Xia, L., Zhao, L., Tan, L., Huang, L., Shi, L., Li, M., Li, M., Cheng, M., Wang, N., Chen, Q., He, Q., Liang, Q., Sun, Q., Sun, R., Wang, R., Pang, S., Yang, S., Liu, S., Liu, S., Gao, S., Cao, T., Wang, T., Ming, W., He, W., Zhao, X., Zhang, X., Zeng, X., Liu, X., Yang, X., Dai, Y., Yu, Y., Li, Y., Deng, Y., Wang, Y., Wang, Y., Lu, Y., Chen, Y., Luo, Y., Luo, Y., Yin, Y., Feng, Y., Yang, Y., Tang, Z., Zhang, Z., Yang, Z., Jiao, B., Chen, J., Li, J., Zhou, S., Zhang, X., Zhang, X., Zhu, Y., Shum, H.Y., Jiang, D.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model (2025), https://arxiv.org/abs/2502.10248
21. Maduabuchi, C.: Entropy-controlled flow matching (2026), https://arxiv.org/abs/2602.22265
22. Maduabuchi, C., Chen, H., Han, Y., Wang, J.: Corruption-aware training of latent video diffusion models for robust text-to-video generation (2026), https://arxiv.org/abs/2505.21545
23. Maduabuchi, C., Wang, J.: Temporal pair consistency for variance-reduced flow matching (2026), https://arxiv.org/abs/2602.04908

24. Masala, M., Cudlenco, N., Rebedea, T., Leordeanu, M.: Explaining vision and language through graphs of events in space and time. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 2818–2823 (2023). https://doi.org/10.1109/ICCVW60793.2023.00302
25. Munir, M., Goel, H., Wei, X., Choi, M., Shah, S., Bhardwaj, K., Whatmough, P., Chinchali, S., Marculescu, R.: Objectalign: Neuro-symbolic object consistency verification and correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3458–3468 (June 2026)
26. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers . In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4172–4182. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00387, https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00387
27. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
28. Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., BAI, L., Ouyang, W., Zhang, R.: Worldsimbench: Towards video generation models as world simulators (2025), https://openreview.net/forum?id=ejGAytoWoe
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042
30. Sharan, S.P., Choi, M., Shah, S., Goel, H., Omama, M., Chinchali, S.: Neuro-Symbolic Evaluation of Text-to-Video Models using Formal Verification . In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8395–8405. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). https://doi.org/10.1109/CVPR52734.2025.00786, https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00786
31. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8406–8416 (2025). https://doi.org/10.1109/CVPR52734.2025.00787
32. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), https://arxiv.org/abs/2503.20314
33. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., Linus, Patrol, Zhang, P., Chen, P., Zhao, P., Tian, Q., Liu, S., Kong, W., Wang, W., He, X., Li, X., Deng, X., Zhe, X., Li, Y., Long, Y., Peng, Y., Wu, Y., Liu, Y., Wang, Z., Dai, Z., Peng, B., Li, C., Gong, G., Xiao, G., Tian, J.,

Lin, J., Liu, J., Zhang, J., Lian, J., Pan, K., Wang, L., Niu, L., Chen, M., Chen, M., Zheng, M., Yang, M., Hu, Q., Yang, Q., Xiao, Q., Wu, R., Xu, R., Yuan, R., Sang, S., Huang, S., Gong, S., Huang, S., Guo, W., Yuan, X., Chen, X., Hu, X., Sun, W., Wu, X., Ren, X., Yuan, X., Mi, X., Zhang, Y., Sun, Y., Lu, Y., Li, Y., Huang, Y., Tang, Y., Li, Y., Deng, Y., Zhou, Y., Hu, Z., Liu, Z., Yang, Z., Yang, Z., Lu, Z., Zhou, Z., Zhong, Z.: Hunyuanvideo 1.5 technical report (2025), https://arxiv.org/abs/2511.18870

34. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=LQzN6TRFg9
35. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3813–3824 (2023). https://doi.org/10.1109/ICCV51070.2023.00355
36. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (October 2023)
37. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness (2025), https://arxiv.org/abs/2503.21755
38. Zheng, K., Lu, C., Chen, J., Zhu, J.: DPM-solver-v3: Improved diffusion ODE solver with empirical model statistics. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=9fWKExmKa0
39. 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 (2024), https://arxiv.org/abs/2412.20404