TunerDiT: Training-free Progressive Steering of Diffusion Transformer for Multi-Event Video Generation

Anonymous authors
Paper under double-blind review

Abstract

Text-to-video (T2V) generation faces challenging questions when generating videos with long horizons containing multiple events. Inspired by the intrinsics of the diffusion process, we probe video diffusion transformers (DiTs) and uncover intrinsic turning points in the DiT denoising trajectory where conditioning text affects generation from global layout to fine-grained details. Building on this finding, we present TunerDiT, a simple yet effective progressive steering method that requires no additional training for multi-event generation. TunerDiT comprises two steering handles: (1) Event-Partitioned Masking that enforces event boundaries while allowing cross-event transition bands; (2) Cross-Event Prompt Fusion that injects neighboring event semantics for late-stage refinement. We contribute a self-curated prompt suite for benchmarking multi-event generation, i.e. *MEve*. TunerDiT achieves the highest text alignment while maintaining visual consistency, and offers a tunable trade-off between video consistency and event separation, compared with other training-free methods. Besides, TunerDiT is model agnostic and the performance is consistent across different model architectures.

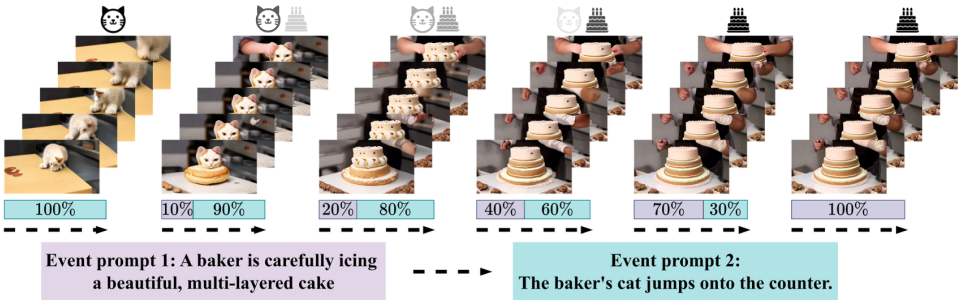


Figure 1: **Different denoising steps use text inputs differently.** TunerDiT derives this insight by probing the prompt conditioning of video diffusion models. When the input prompt is switched at various fractions of the denoising steps (e.g., 0%, 10%, 20%, 40%, 70%, 100%), early steps dominate the global layout while late steps refine fine-grained appearance and motion, revealing intrinsic turning points where text influence changes. Building on this insight, TunerDiT is a training-free, progressive steering method for video DiTs that yields consistent multi-event videos with clear boundaries and smooth transitions.

1 Introduction

Recent advances in text-to-video (T2V) generation have been largely driven by diffusion transformers (DiTs) (Peebles & Xie, 2023; Chen et al., 2024b; Xing et al., 2024) . Models such as CogVideoX (Yang et al., 2024) and Open-Sora (Zheng et al., 2024) adopt DiT-style architectures with spatial-temporal self-attention, 3D VAEs (Li et al., 2024), and large-scale pretraining, achieving impressive visual quality on short clips for one good moment, which are insufficient for long-horizon, multi-event scenarios required by many real-world applications (Li et al., 2025; Kang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2024; Cho et al., 2024; Liu et al., 2025; Yang et al., 2025; Liao et al., 2024a;b), such as long-horizon robotic planning, autonomous driving, interactive world models, etc.

When naively repurposed for multi-event generation via direct concatenation of event prompts in time or text, existing methods exhibit clear, distinct failure patterns. The resulting videos blend all events into a single muddled scene. The core desiderata of multi-event video generation are: (1) *Event ordering*: clear event boundaries and correct temporal ordering; (2) *Smooth transitions*: smooth handovers between events; (3) *Semantic Consistency*: fine-grained semantic consistency across events bidirectionally between early and late events.

Only limited work directly addresses multi-event T2V generation (Cai et al., 2025; Oh et al., 2024; Wang et al., 2023; Qi et al., 2025; Qiu et al., 2024; Liao et al., 2025; Xu et al., 2026). Achieving these requirements simultaneously has been proven computationally expensive (Wu et al., 2025; Kara et al., 2025) and hard to scale to more events beyond the training scope. Other computationally light works achieve adequate but not satisfactory performance (Wang et al., 2023; Cai et al., 2025; Oh et al., 2024). However, the denoising dynamics and cross-attention structure of DiTs point to a principled, tractable approach that meets all these demands.

Inspired by observations in text-to-image (T2I) diffusion that generation proceeds from coarse global structure to fine-grained details over the denoising trajectory (Balaji et al., 2022; Chen et al., 2024a; Yang et al., 2023), we hypothesize that Video DiTs exhibit analogous behavior in T2V generation, given the shared intrinsic properties of the diffusion process. Under this assumption, the temporal ordering of video events can be viewed as an additional dimension of fine-grained detail to be refined over the denoising process.

To validate the hypothesis, we first curate a benchmark suite *MEvE* that consists of multi-event prompts from multiple aspects. After systematically probing off-the-shelf DiT video generators by switching event conditioning at different denoising timestep fractions, we observe the resulting text-video similarity (Kara et al., 2025) as shown in Figure 1. Consequently, we identify that text conditioning on early steps dominates the global video layout while text conditioning on late steps refines fine-grained appearance and motion, revealing the intrinsic turning point where text influence changes and is remarkably stable within a model.

Building on this insight, we introduce **TunerDiT**, a progressive coarse-to-fine steering framework for multi-event T2V generation that operates *without training*. The idea is to first generate a shared layout of multiple events and refine inner event details at a certain later stage. TunerDiT exploits the *intrinsic turning points* of the DiT denoising process through two steering handles activated on a schedule: **Cross-Event Prompt Fusion** is a gating scheme that conditions video latents on event prompts to enhance semantic awareness and coherence, and **Event-Partitioned Mask** is a diagonal mask on DiT attention layers that isolates events and links neighbors through transition bands for clear boundaries and smooth handovers.

Extensive experiments validate the effectiveness of TunerDiT. We achieve state-of-the-art results in multi-event video generation up to four events, with principled extensibility to more. TunerDiT consistently improves text–video alignment over strong baselines while maintaining background and identity consistency. TunerDiT runs at inference time on a single A100 GPU, making it cost-effective for real-world deployment. Our contributions are:

- We propose **MEvE**, a multi-aspect benchmark prompt suite with up to 4 events per prompt.
- We reveal the **intrinsic turning point** from coarse to fine-grained generation in Video DiTs.
- We propose TunerDiT, a tuner console-like steering approach for video diffusion transformers to generate multi-event consistent videos under a training-free setting.

2 Related Work

DiT-based Video Generation Recent works such as CogVideoX (Yang et al., 2024), Mochi 1 (Team, 2024), Movie Gen (Polyak et al., 2024), HunyuanVideo (Kong et al., 2024), and Wan (Team Wan et al., 2025) have adopted the DiT (Peebles & Xie, 2023) architecture and achieved impressive results on generating one good moment. In this work, we build upon the open-source Open-Sora (Zheng et al., 2024; 2025b) and Wan models. The Open-Sora architecture evolved from a Spatial-Temporal DiT (ST-DiT) structure to a dual-stream DiT. ST-DiT employs separate spatial and temporal attention

blocks for frame-level and temporal processing, while the dual-stream DiT integrates image and text through parallel streams with one unified attention matrix.

Multi Event Video Generation Autoregressive methods extend videos from previous frames but lose global control over multiple events (Kim et al., 2024; Wang et al., 2023). Training-based methods modify DiT attention with event-scoped temporal cross-attention (Wu et al., 2025), cross-event transition tokens (Kara et al., 2025), or symmetric text–video masks that isolate events at the cost of sharp cuts (Qi et al., 2025), all requiring intensive computation and labeled video data. Zero-shot methods initialize latents from the last frame of the previous event (Oh et al., 2024), impose compositional layout priors without temporal order (Zhang et al., 2025), share subject attention through segmentation masks (Cai et al., 2025), or steer frame-level queries toward their aligned event prompts (Xu et al., 2026). TunerDiT is the first zero-shot method to address event ordering, smooth transitions, and global semantic consistency together, through progressive steering grounded in the intrinsics of the diffusion process.

3 The Intrinsic Turning Point in DiT

In this section, we first introduce the failure mode of general DiT-based Video Diffusion Models in generating multi-events, then introduce the hypothesis of the varying text conditioning effects at different diffusion step fractions in DiTs. To validate this hypothesis, we propose a new benchmark MEve and probe the intrinsic points in the DiT models, and show the resulting insight.

3.1 Failure Modes in DiT-based Models

Existing DiT video diffusion models exhibit three failure modes on multi-event prompts (Figure 2). **(a) Event fusion:** all prompts are entangled into a single blended scene, and objects and actions from different events co-occur in most frames. **(b) Scrambled order:** events appear in random order, overlap with earlier ones, or disappear entirely. **(c) Transition collapse:** transitions are abrupt frame-wise jumps or overly smooth morphs without clear event boundaries.

Open-Sora 2.0 on multi-event prompts	Event 1	Event 2	Event 3
	…one wearing a **red** shirt on the left and the other wearing a **blue** shirt.	…focus on the speaking person wearing a **red** shirt directly.	…focus on the speaking person wearing a **blue** shirt directly.
	A man is walking in the park.	Then the man sits on the bench and starts smoking.	The man then throws the cigarette in the trash.
	A bird is flying over a yellow grassland.	The bird then flies across a forest.	The bird finally lands on a rock beside a lake.

Figure 2: Open-Sora 2.0 fails on multi-event prompts: each row shows sampled frames generated from the three event prompts on its right.

3.2 Benchmark, Hypothesis & Validation: Intrinsic Turning Point

Hypothesis: from T2I to T2V. Text-to-image (T2I) diffusion generates images from coarse to fine (Balaji et al., 2022; Chen et al., 2024a; Yang et al., 2023). We hypothesize that Video DiTs share this behavior in text-to-video (T2V) generation, so the succession of video events can be treated as an additional dimension of fine-grained detail rather than generated at once.

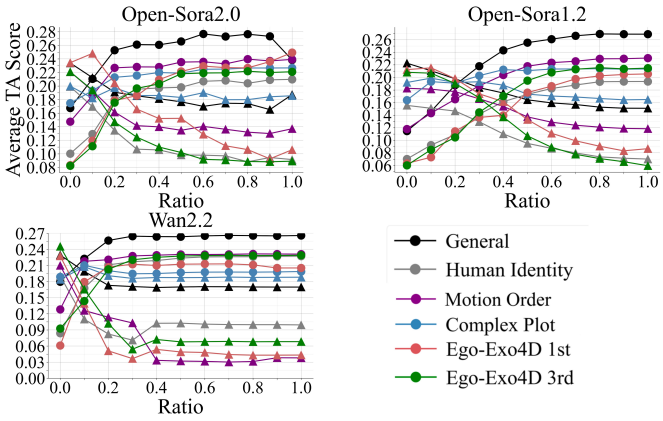
Validation: Switching Denoising Fractions. To test this hypothesis, we condition a DiT on a first prompt P_1 for a fraction x of the denoising steps and switch to a second prompt P_2 for the rest (Figure 1). Even $x = 10\%$ imprints P_1 on the layout: the cake of P_1 already appears, while P_2 contributes only cat-like appearance. At $x = 20\%$ the cake dominates and the cat survives only as decoration, and from $x = 40\%$ on the video follows P_1 almost entirely. Early steps therefore fix the global layout, late steps only refine appearance and motion, and control passes from one prompt to the other at an *intrinsic turning point* $\tau^\star$ of the denoising trajectory. This phenomenon offers a

Category	View	Events	#Prompts
General	3rd	4	60
Motion Order	3rd	4	98
Human Identity	3rd	4	32
Complex Plot	3rd	4	60
Ego-Exo4D (paired)	1st/3rd	4	50×2

(a) MEve composition.

Benchmark	Shot Counts	Multi-Scene	Multi-Event	Real Ref.	View Control
VBench	1	×	×	×	×
VBench++	1	×	×	×	×
VBench2.0	1	×	×	×	×
WorldScore	3	✓	×	✓	✓
MPVBench	2	✓	×	×	✓
MEve	4	✓	✓	✓	✓

(b) Comparison with benchmarks.



(c) Turning-point probing on MEve.

Figure 3: MEve and turning-point analysis. (a) MEve composition per category. (b) Position of MEve among existing benchmarks. (c) Probing results revealing intrinsic turning points.

lever for multi-event generation: a prompt shared during the early steps can fix one scene layout for all events, and event-specific prompts applied after the turning point can shape each event without breaking it. We therefore build a benchmark to locate the turning point and design TunerDiT to steer around it, exploiting how the denoiser uses text differently at different noise levels.

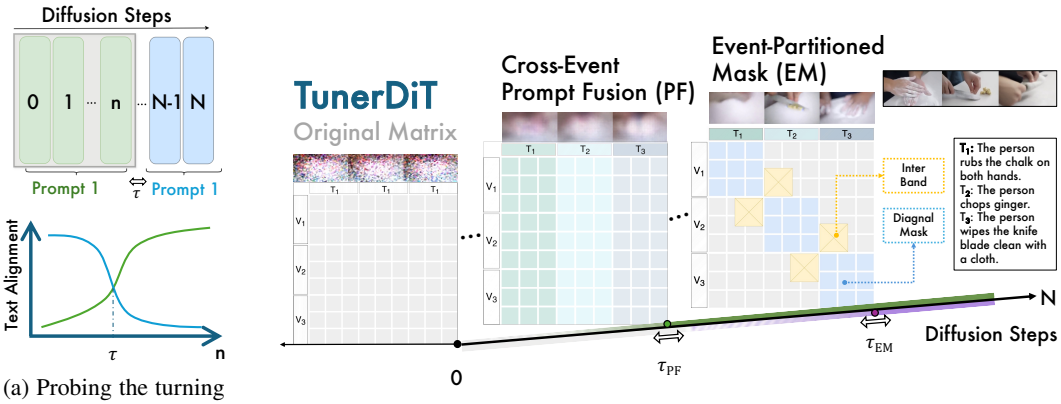
Benchmark: MEve Existing T2V benchmarks focus on single-shot generation or dual-shot transitions and seldom cover multi-event prompts (Huang et al., 2024a; Zheng et al., 2025a; Huang et al., 2024b). We therefore construct MEve, a suite of prompts with up to four events from three complementary sources: (1) LLM-synthesized narratives (Kara et al., 2025); (2) diagnostic prompts expanded from VBench 2.0 categories (Zheng et al., 2025a; Zhang et al., 2024); and (3) paired first- and third-view prompts curated from Ego-Exo4D narrations (Grauman et al., 2024). Construction details and statistics are in Appendix A, Figure 3a, and Appendix Figure 7.

Probing: from Coarse to Fine-grained. To locate $\tau^\star$ on MEve (Figure 4a), we formalize the switching schedule as $\text{cond}(n) = \text{P}_1$ for denoising step $n < xN$ and P_2 otherwise, and sweep the fusion ratio x over $\mathcal{X} = \{0, 0.1, \ldots, 1.0\}$. We set $\tau^\star = \lfloor xN \rfloor$ at the latest x where the Text Alignment (TA) curves of P_1 and P_2 intersect, i.e., the step where the DiT hesitates between the two event layouts in deciding which event dominates the global scene composition and camera layout.

The position of the intrinsic turning point. Across models and event categories (Figure 3c), TA shifts sharply within the early steps: the dominance of the second prompt changes at $x \in [0, 0.3]$, followed by a broad plateau. Exposing the model to a new event within the first ∼30% of steps thus suffices to trigger a shift, while later steps have diminishing influence. $\tau^\star$ thus marks the phase boundary between layout planning and detail refinement, which TunerDiT uses as a controllable "tuning knob" for multi-event generation: steering before $\tau^\star$ reshapes the shared layout, while steering after it refines the appearance and motion of each event without altering that layout. Since $\tau^\star$ sits in an approximate region in the denoising process, it is determined empirically as a hyperparameter that this paper later utilize an interval for searching it.

4 Method: TunerDiT

TunerDiT (Figure 4, Alg. 1) comprises two steering handles: an *Event-Partitioned Mask* (EM, §4.1) that separates events and links them smoothly, and *Cross-Event Prompt Fusion* (PF, §4.2) that builds a shared global layout. A gating schedule (§4.3) activates at the *intrinsic turning point* (§3).



(a) Probing the turning point.

(b) Turning point guided steering controls applied over denoising steps.

Figure 4: **TunerDiT** progressively steers multi-event generation over diffusion steps. *Cross-Event Prompt Fusion (PF)* first shares a common prompt to build a coherent layout, then separates event prompts. *Event-Partitioned Mask (EM)* subsequently enforces event isolation via diagonal cross-attention blocks and introduces cross-event transition bands around boundaries, enabling smooth and semantically consistent handovers between events.

4.1 Event-Partitioned Diagonal Mask

Let E be the number of events and P_e the prompt of event e with text embeddings $T_e = f_{\text{text}}(P_e) \in \mathbb{R}^{m_e \times d}$. The *Event-Partitioned Diagonal Mask* (EM) separates events while preserving smooth transitions. We split the video tokens into E consecutive segments $V = [V_1; \ldots; V_E]$ and the text tokens into the matching prompts $T(n) = [T_1(n); \ldots; T_E(n)]$, where V_e and $T_e(n)$ belong to event e.

Algorithm 1 Progressive steering of TunerDiT

Require Event embeddings $\{T_1, \ldots, T_E\}$, video embedding V, total denoising steps N
Require Candidate ratios $x_{\text{EM}}^{\text{list}}, x_{\text{PF}}^{\text{list}}$ around $\tau^\star/N$; band ratio r
Ensure DiT attention steered by PF and EM

▷ **EM**: select the best event-partitioning ratio
$x_{\text{EM}} \leftarrow \arg\max_{x \in x_{\text{EM}}^{\text{list}}} \text{TA}_{\text{EM}}(x)$
$\tau_{\text{EM}} \leftarrow \lfloor N \cdot x_{\text{EM}} \rfloor$
▷ ***PF**: select the best prompt-fusion ratio*
$x_{\text{PF}} \leftarrow \arg\max_{x \in x_{\text{PF}}^{\text{list}}} \text{TA}_{\text{PF|EM}}(x)$
$\tau_{\text{PF}} \leftarrow \lfloor N \cdot x_{\text{PF}} \rfloor$
for $n = 0$ **to** $N-1$ **do**
  ▷ **PF**: activate prompt fusion
  if $n \le \tau_{\text{PF}}$ **then**
    $T(n) \leftarrow \text{concat}(T_1, T_1, \ldots, T_1)$
  else
    $T(n) \leftarrow \text{concat}(T_1, T_2, \ldots, T_E)$
  end if
  ▷ **EM**: activate event mask
  if $n \ge \tau_{\text{EM}}$ **then**
    $r(n) \leftarrow \text{ReLU}\big(r - (n - \tau_{\text{EM}})/N\big)$
    $D(n) \leftarrow \text{GetEventMask}(E, V, r(n))$
  else
    $D(n) \leftarrow \varnothing$
  end if
  ▷ **DiT**: steered update
  $V \leftarrow \text{DiTAttention}(V, T(n), D(n))$
end for

Event-partitioned diagonal mask. The diagonal mask keeps each event segment focused on its own event prompt. Specifically, for video tokens in segment V_e, EM mainly allows attention to the matching text tokens $T_e(n)$ and suppresses attention to unrelated event prompts. This produces the event-wise diagonal pattern $D_{\text{diag}} = \text{diag}(V_1 \leftrightarrow T_1, \ldots, V_E \leftrightarrow T_E)$, the binary-pass attention map for event-matched video–text pairs. This prevents cross-event semantic leakage and preserves event order.

Transition bands. Transition bands allow neighboring events to communicate only near their boundary. A purely diagonal mask can create hard cuts between adjacent events, so for each boundary $e \to e+1$, we augment the diagonal structure with narrow bands. These *inter-event* bands $D_{\text{band}} = D_{\text{inter}}^{v \leftarrow t} \vee D_{\text{inter}}^{t \leftarrow v}$ let boundary video tokens attend to the neighboring event's text tokens in video–text cross-attention. A ratio $r \in (0, 1)$ sets the band widths $h_e^{vt} = \lfloor r|V_e| \rfloor$ and $w_e^{vt} = \lfloor r|T_e| \rfloor$ for event e with $|V_e|$ video and $|T_e|$ text tokens: h_e^{vt} determines the number of video tokens near the boundary that participate in the transition, and w_e^{vt} determines the number of neighboring text tokens visible in cross-attention.

4.2 Cross-Event Prompt Fusion

Cross-Event Prompt Fusion (PF) decides which event prompt each video segment is conditioned on. PF first *fuses* the events by conditioning every segment on the first-event prompt, so that all segments share one global scene and camera layout, and then *unfuses* them by handing each segment its own event prompt. Concretely, PF replaces the text tokens with effective tokens $\widehat{T}_e(n)$ and $\widehat{T}(n) = [\widehat{T}_1(n); \ldots; \widehat{T}_E(n)]$, scheduled over diffusion steps as defined in §4.3.

4.3 Temporal Gating

Having probed the intrinsic turning point $\tau^\star$ of a model, TunerDiT sets two tunable gates near $\tau^\star$: τ_{PF} for prompt fusion and τ_{EM} for the event mask (Algorithm 1). Steps before the gates build a coherent global layout; steps after them make the DiT distinguish events. The turning point is a relative position that differs across DiTs, so both gates are per-model hyperparameters whose candidate grids in Algorithm 1 lie in a window around $\tau^\star/N$ (Appendix B).

Gating of PF. For diffusion step $n \in \{1, \ldots, N\}$, the effective text tokens are

$$\widehat{T}_e(n) = T_1 \text{ if } n \le \tau_{\text{PF}}, \text{ else } T_e. \tag{1}$$

Thus, in the *coarse layout diffusion phase* when $n \le \tau_{\text{PF}}$, all video tokens are conditioned on the first-event prompt P_1, encouraging a globally consistent scene and camera layout shared across the whole sequence. After the gate τ_{PF}, the model starts to differentiate between events content details by conditioning video embeddings on their own event text embeddings, effectively "unfusing'' the prompts and handing control over to event-specific conditioning.

Gating of EM. EM is disabled before τ_{EM} and enabled only after the global layout has stabilized. Before τ_{EM}, all video tokens can attend to all text tokens. After τ_{EM}, only the event-wise diagonal blocks and transition bands are allowed. The transition bands follow a ReLU-shaped schedule: they stay inactive until τ_{EM}, then start at the full ratio r and shrink linearly to zero with unit slope along the normalized diffusion time:

$$D_{\text{EM}}^{v\leftarrow t}(n) = \begin{cases} \mathbf{1}, & n < \tau_{\text{EM}}, \\ D_{\text{diag}} \vee D_{\text{band}}(n), & n \ge \tau_{\text{EM}}, \end{cases}$$

$$r(n) = \begin{cases} 0, & n < \tau_{\text{EM}}, \\ \text{ReLU}\big(r - \frac{n-\tau_{\text{EM}}}{N}\big), & n \ge \tau_{\text{EM}}, \end{cases}$$

where $\vee$ is element-wise OR and $D_{\text{band}}(n)$ uses the band widths of §4.1 with $r(n)$ in place of r. The bands vanish at step $\tau_{\text{EM}} + rN$, after which only the diagonal blocks remain. The additive mask $M_{\text{EM}}^{v\leftarrow t}(n) = (-\infty) \cdot (\mathbf{1} - D_{\text{EM}}^{v\leftarrow t}(n))$ is added to the attention logits before the softmax. For text-to-video attention, we use the transpose of the same mask. For video self-attention, we apply the same event-wise blocking rule without bands. Thus, EM isolates unrelated events while preserving enough boundary context for gradual transitions.

5 Experiments

5.1 Experimental Setup

Baselines To comprehensively evaluate the proposed framework, we benchmarked it against a diverse set of recent text-to-video methods covering both single-event and multi-event generation paradigms. For open-source T2V large models, we included Open-Sora 1.2/2.0 (Zheng et al., 2024; 2025b) and Wan2.1/2.2 (Team Wan et al., 2025) as strong diffusion-transformer backbones representing large-scale open models. We further compared with Training-free baselines for fair comparison, i.e., MEVG (Oh et al., 2024), DiTCtrl (Cai et al., 2025), FreeNoise (Qiu et al., 2024), and SwitchCraft (Xu et al., 2026). We also report each base video generation model without steering to isolate the gain contributed by TunerDiT. Implementation details are in Appendix B.

Table 1: Quantitative comparison. (a) Automatic metrics over 2–4 events. (b) VLM-judge and human-study metrics, with the best value in each column in **bold** and the second best underlined.

(a) Comparison with training-free methods and with each base model before and after (↪) TunerDiT. **Bold**: TA and TIS values that improve over the base model. On average, TunerDiT raises TA by 11.5% and TIS by 2.5%, while BC, IC, and MS change by only −0.6%.

	Text alignment (improved)						*Consistency (maintained)*								
	TA↑			**TIS↑**			**BC↑**			**IC↑**			**MS↑**		
Shot Number	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**
Training-free Multi-Event Methods															
MEVG	0.201	0.206	0.205	0.271	0.270	0.272	0.839	0.842	0.821	0.711	0.718	0.709	0.941	0.932	0.928
DiTCtrl	0.186	0.207	0.216	0.259	0.271	0.278	0.872	0.888	0.881	0.767	0.765	0.770	0.972	0.979	0.976
FreeNoise	0.197	0.199	0.206	0.272	0.267	0.275	0.863	0.870	0.854	0.725	0.747	0.709	0.962	0.967	0.958
SwitchCraft (Wan2.1)	0.182	0.194	0.205	0.265	0.266	0.275	0.943	0.928	0.935	0.924	0.891	0.903	0.976	0.974	0.972
Base Video Generation Models															
Open-Sora 1.2	0.174	0.179	0.182	0.270	0.269	0.276	0.941	0.940	0.945	0.904	0.908	0.901	0.985	0.983	0.988
↪ + TunerDiT	**0.202**	**0.211**	**0.213**	**0.277**	**0.284**	**0.281**	0.943	0.938	0.937	0.899	0.891	0.894	0.981	0.984	0.983
Open-Sora 2.0	0.194	0.198	0.199	0.270	0.276	0.281	0.929	0.932	0.931	0.871	0.867	0.861	0.991	0.990	0.984
↪ + TunerDiT	**0.210**	**0.211**	**0.219**	**0.280**	**0.277**	**0.287**	0.924	0.928	0.927	0.848	0.851	0.850	0.981	0.982	0.984
Wan2.1	0.182	0.184	0.183	0.265	0.267	0.263	0.944	0.945	0.947	0.922	0.918	0.919	0.979	0.977	0.981
↪ + TunerDiT	**0.210**	**0.212**	**0.209**	**0.275**	**0.272**	**0.271**	0.940	0.939	0.937	0.918	0.910	0.913	0.977	0.981	0.980
Wan2.2	0.185	0.201	0.201	0.270	0.271	0.275	0.921	0.916	0.919	0.849	0.847	0.851	0.987	0.985	0.983
↪ + TunerDiT	**0.201**	**0.208**	**0.211**	**0.273**	**0.277**	**0.279**	0.913	0.914	0.916	0.844	0.850	0.848	0.981	0.982	0.984

(b) Left: VLM-judge Event Isolation (EI) and Text–Video Alignment (TVA). Right: human study (18 people).

	EI			**TVA**		
	2	**3**	**4**	**2**	**3**	**4**
Training-free Methods						
MEVG	0.360	0.435	0.515	1.335	1.375	1.451
FreeNoise	0.362	0.436	0.469	1.360	1.400	1.514
DiTCtrl	0.305	0.375	0.449	1.385	1.425	1.479
SwitchCraft (Wan2.1)	0.213	0.275	0.342	1.387	1.599	1.618
Ours						
TunerDiT Open-Sora 1.2	0.423	0.499	0.580	1.450	1.492	1.524
TunerDiT Open-Sora 2.0	**0.470**	**0.572**	**0.660**	**1.490**	1.533	1.631
TunerDiT Wan2.1	0.419	0.541	0.628	1.437	**1.619**	**1.648**
TunerDiT Wan2.2	0.399	0.474	0.553	1.462	1.503	1.589

	Human rating			
	Q1	**Q2**	**Q3**	**Q4**
Training-free Methods				
MEVG	1.82	1.91	1.82	2.05
FreeNoise	1.99	2.01	1.99	1.79
DiTCtrl	2.11	2.03	2.18	2.35
SwitchCraft (Wan2.1)	2.78	2.80	2.98	2.67
Ours				
TunerDiT Open-Sora 1.2	2.66	2.63	2.62	2.94
TunerDiT Open-Sora 2.0	3.16	3.03	**3.03**	3.34
TunerDiT Wan2.1	**3.25**	**3.20**	3.00	**3.40**
TunerDiT Wan2.2	3.01	2.97	2.82	3.30

Metrics We evaluate multi-event generation from three complementary perspectives: (1) automatic embedding/segmentation-based metrics adapted from single-shot T2V protocols; (2) VLM-as-judge scores; and (3) human evaluations.

(1) Embedding-based automatic metrics. Following the T2V protocols of Wu et al. (2025) and their multi-shot adaptation in Kara et al. (2025), we report five metrics. **Text Alignment (TA)** is the cosine similarity between event-level text features and video features from ViCLIP (Wang et al., 2024), averaged across events. **Text–Image Similarity (TIS)** is the CLIP similarity (Radford et al., 2021) between the prompt and uniformly sampled frames, following VBench 2.0 (Zheng et al., 2025a), and reflects global alignment of the full clip. The other three are VBench (Huang et al., 2024a) dimensions: **Background Consistency (BC)** measures the CLIP feature similarity of the background across frames; **Identity Consistency (IC)**, VBench's subject consistency, measures the DINO feature similarity of the main subject across frames; and **Motion Smoothness (MS)** checks generated motion against a frame-interpolation motion prior.

(2) VLM-as-judge metrics. We uniformly sample frames of each video and query Gemini-2.5-Flash (Comanici et al., 2025) for per-frame captions and a clip summary. **Event Isolation (EI)** is the accuracy of the event count that the VLM predicts from the summary against the ground-truth count. **Text–Video Alignment (TVA)** is the Sentence-BERT (Reimers & Gurevych, 2019) cosine similarity between the multi-event prompt and the VLM summary, averaged over videos.

(3) Human evaluation. Finally, we conduct a user study of 18 persons on a 5-point Likert scale, where annotators compare different models on four aspects: (Q1) overall video preference, (Q2) motion naturalness, (Q3) transition smoothness across events, and (Q4) text–video alignment. Detailed protocol, interface, annotator instructions, and additional statistics are provided in Appendix C.

5.2 Experiment Results: Quantitative Analysis

Multi-event steering quality. ***Higher text alignment with maintained consistency.*** On all four backbones and at every event count, TunerDiT improves TA and TIS over its base model and attains

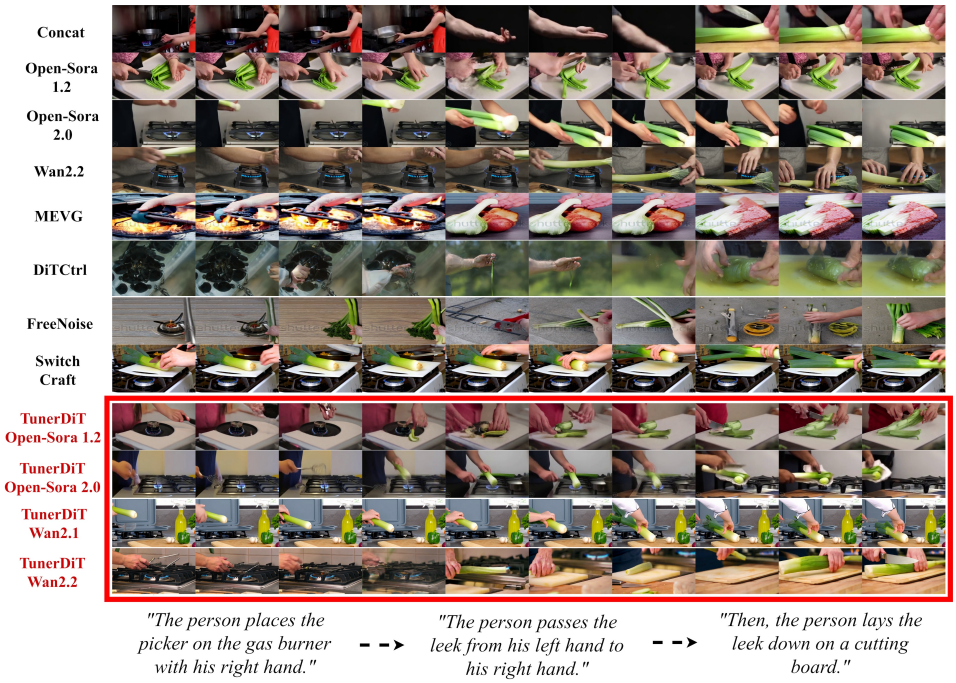


Figure 5: Qualitative comparison to various open-source base models (Open-Sora 1.2, Open-Sora 2.0, Wan2.1, and Wan2.2), direct concatenation of individual events (Concat), and training-free multi-event baselines (MEVG, DiTCtrl, FreeNoise, and SwitchCraft). Our methods achieve the best performance in event ordering, transition smoothness, and semantic-aware consistency.

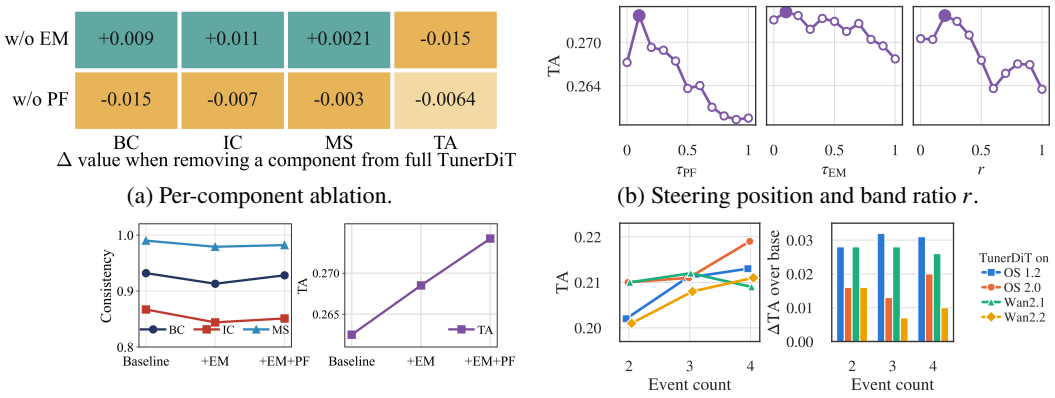
the highest TA and TIS among all methods, while keeping BC, IC, and MS close to the base model (Table 1a). On average, TA rises by 11.5% (up to 17.9%) and TIS by 2.5%, whereas BC, IC, and MS change by only −0.6% (at most −2.6%): the TA gain is about 20× and the TIS gain about 4× the consistency cost. Its TA rises with the event count on three of four backbones (Figure 6d). ***Best human-aligned preference.*** TunerDiT variants attain the best VLM-judge score in every EI and TVA column and the top human rating on every question, led by TunerDiT (Wan2.1) on overall preference, motion naturalness, and text–video alignment, exceeding the strongest baseline SwitchCraft by up to +0.73 (Table 1b). ***Stronger on a shared backbone.*** On Wan2.1, TunerDiT surpasses SwitchCraft on TA, MS, EI, and TVA at every event count and on every human-study question (Table 1), an advantage that holds against a competing method on an identical backbone.

5.3 Ablation Study

Figures 6a and 6c illustrate how each steering component in TunerDiT contributes to multi-event generation quality. *Text alignment* (TA, TIS) and *visual consistency* (BC, IC, MS) are in tension: improving one group often degrades the other, and TunerDiT balances them.

Effect of Event-Partitioned Mask (EM). As shown in Figure 6a, enabling EM improves multi-event compliance. However, this comes at the cost of lower BC/IC/MS, since boundary enforcement caused by diagonal and transition masks suppresses cross-event leakage, which inevitably harms global smoothness and consistency. The increase of TA/TIS indicates that events become more cleanly separated and better aligned with their prompts.

Recovering consistency with Prompt Fusion (PF). Adding PF in TunerDiT restores a significant portion of the lost consistency while preserving most of the gains in multi-event alignment. In Figure 6c, adding PF on top of EM raises BC (0.727 → 0.794) and IC (0.598 → 0.677) while further improving TA (0.268 → 0.274). PF therefore acts as a counter-balance to EM, yielding a more favorable trade-off between the two mutually inhibiting metric groups: event separation versus semantically coherent across events. More detailed discussion is positioned in Appendix D.



(a) Per-component ablation.

(b) Steering position and band ratio r.

(c) Module combinations.

(d) Improvement with the event count.

Figure 6: Ablation and analysis of TunerDiT. (a) Contribution of each steering component. (b) Steering acts mostly early in diffusion. (c) Module combinations on Open-Sora 2.0. (d) TA of the four TunerDiT variants (left) and their ΔTA over the base model (right) per event count.

5.4 Qualitative Analysis

Figure 5 compares multi-event sequences for a three-step cooking prompt. Concat generates each event separately and concatenates the clips, which yields the worst semantic consistency. The three base models, prompted with all events at once, fail the three challenges and produce the wrong event count and order. DiTCtrl loses text alignment and consistency when events shift motions. TunerDiT achieves the best visual quality in event ordering, transition smoothness, and consistency. Appendix E shows more examples. Additional videos are available in the supplementary material.

5.5 Discussions

Appendix F covers failure cases, event types, limitations, and future work.

Investigation on steering position and transition bandwidth ratio. As shown in Figure 6b, TA is most sensitive to the prompt-fusion gate τ_{PF} and peaks when τ_{PF} lies early ($\approx$0.0–0.2 of the steps), confirming that early layout steering is crucial for multi-event alignment. The EM gate τ_{EM} shows a flat plateau over a broad mid-range that still peaks around 0.1–0.2, before the late refinement steps. Small to moderate transition bands ($r \approx 0.1$–0.3) yield the best TA, while wide bands ($r > 0.6$) over-smooth boundaries until events are no longer distinguished. We therefore adopt an early τ_{PF}, a mid-range τ_{EM}, and a moderate r by default, which jointly give the strongest multi-event alignment.

Model-agnostic gains across event counts. TA rises with the event count on three of the four TunerDiT variants (Figure 6d, left): from 0.202 to 0.213 on Open-Sora 1.2, 0.210 to 0.219 on Open-Sora 2.0, and 0.201 to 0.211 on Wan2.2. TunerDiT is model-agnostic: it raises TA over the base model on all four backbones at every event count, by +0.007 to +0.032 (right). Each added event adds a boundary, which event-partitioned steering turns into an anchor that stabilizes ordering and curbs cross-segment leakage.

6 Conclusion

We presented TunerDiT, a training-free progressive steering framework for multi-event T2V generation that exploits the intrinsic turning point where text guidance in DiTs shifts from global layout to fine details: prompt fusion shares one layout across events before it, and the event-partitioned mask separates events after it with smooth transitions. TunerDiT raises text alignment by 11.5% on average at a consistency change of only −0.6%, attains the best VLM-judge and human preference scores, and surpasses SwitchCraft on a shared backbone.

AI USE STATEMENT

In this work, we used generative AI tools as components of the benchmark and evaluation pipeline: Gemini 2.5 Pro generates the LLM-synthesized multi-event prompts of MEVE from controlled instructions (Appendix A), and Gemini-2.5-Flash serves as the VLM judge that captions sampled frames and summarizes each generated video for the Event Isolation and Text–Video Alignment metrics (Section 5.1 and Appendix G). The text-to-video models that TunerDiT steers and compares against are objects of study, not tools for producing the paper's content. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for improving the manuscript text and its LaTeX layout and for polishing the scripts that draw the figures. We have reviewed all AI-assisted work: the LLM-synthesized prompts were checked by the authors before inclusion in MEVE, the VLM-judge scores are complemented by a human study with 18 participants (Appendix C), and AI-assisted text and figure code were checked, e.g. plotted values were verified against the result tables and raw experiment logs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code will be released for full reproducibility after the review period.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7763–7772, 2025.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5343–5353, 2024a.

Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation, 2024b. URL https://arxiv.org/abs/2312.04557.

Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo,

Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024. URL https://arxiv.org/abs/2311.18259.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024b. URL https://arxiv.org/abs/2411.13503.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28405–28415, 2025.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37:89834–89868, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan. Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. *arXiv preprint arXiv:2411.17459*, 2024.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6577–6602, 2024a.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. Gentkg: Generative forecasting on temporal knowledge graph with large language models. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 4303–4317, 2024b.

Ruotong Liao, Guowen Huang, Qing Cheng, Thomas Seidl, Daniel Cremers, and Volker Tresp. When and where do events switch in multi-event video generation?, 2025. URL https://arxiv.org/abs/2510.03049.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception. *arXiv preprint arXiv:2508.15720*, 2025.

Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. Mevg: Multi-event video generation with text-to-video models. In *European Conference on Computer Vision*, pp. 401–418. Springer, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Mask^ 2dit: Dual mask-based diffusion transformer for multi-scene long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18837–18846, 2025.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2024. URL https://arxiv.org/abs/2310.15169.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Volker Tresp, Sahand Sharifzadeh, Hang Li, Dario Konopatzki, and Yunpu Ma. The tensor brain: A unified theory of perception, memory, and semantic decoding. *Neural Computation*, 35(2): 156–227, 01 2023. ISSN 0899-7667. doi: 10.1162/neco_a_01552. URL https://doi.org/10.1162/neco_a_01552.

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2307.06942.

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23989–24000, 2025.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models, 2024. URL https://arxiv.org/abs/2310.10647.

Qianxun Xu, Chenxi Song, Yujun Cai, and Chi Zhang. Switchcraft: Training-free multi-event video generation with attention controls, 2026. URL https://arxiv.org/abs/2602.23956.

Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 36:26291–26303, 2023.

Zhuoran Yang, Xi Guo, Chenjing Ding, Chiyu Wang, Wei Wu, and Yanyong Zhang. Instadrive: Instance-aware driving world models for realistic and consistent video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25410–25420, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv preprint arXiv:2412.09645*, 2024.

Hongyu Zhang, Yufan Deng, Shenghai Yuan, Peng Jin, Zesen Cheng, Yian Zhao, Chang Liu, and Jie Chen. Magiccomp: Training-free dual-phase refinement for compositional video generation. *arXiv preprint arXiv:2503.14428*, 2025.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025a.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, et al. Open-sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025b.

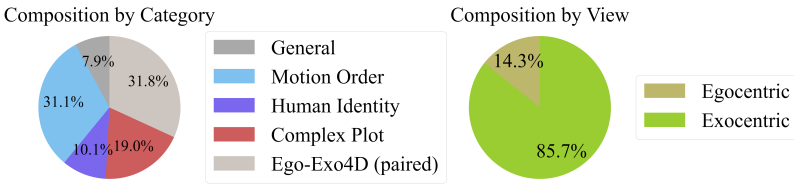


Figure 7: Category and view distribution of MEve.

A MEve: Benchmark Construction Details

Existing Multi Prompt Benchmarks We compare a large body of recent work focusing on building *general-purpose* video generation benchmarks as shown in Figure 3b. VBench (Huang et al., 2024a) decomposes "video generation quality'' into 16 hierarchical dimensions. VBench++ (Huang et al., 2024b) further extends this line by supporting image-to-video settings by introducing an image suite. VBench-2.0 (Zheng et al., 2025a) shifts the focus from superficial visual plausibility to intrinsic faithfulness. These benchmarks diagnose *overall* model quality within a single semantic unit.

For *multi-prompt* video generation, ShotAdapter (Kara et al., 2025) uses a general ChatGPT curated dataset, which lacks discussion for different prompt types. DiTCtrl (Cai et al., 2025) introduces MPVBench, a dedicated benchmark for evaluating multi-prompt transitions with ten transition types and represents an important first step toward benchmarking multi-prompt video generation. However, MPVBench primarily focuses on evaluating the transition effects between two similar scenes that do not differ significantly in motion content. It deliberately avoids logical sequential prompts; hence, this benchmark is not generalizable to real-world scenarios where temporal and logical events occur. We follow the definition of a single event unit that consists of four basic components, *i.e., subject, predicate, objects, and time (Tresp et al., 2023)*, and prior benchmarks do not explicitly address this definition and have a mixture of multi-scene with multi-event settings, such as WorldScore (Duan et al., 2025), which lacks the four event components.

Motivation and MEve Construction Details In contrast, our goal with MEve is to provide a targeted benchmark for *multi-event* video generation, built from three complementary sources: (1) conventional LLM-generated prompts tailored to multi-event stories; (2) multi-event variants of existing general video benchmarks (e.g., VBench 2.0 (Zheng et al., 2025a)) obtained by transforming single-scene prompts into multi-event counterparts; and (3) real-world event sequences derived from egocentric videos with natural-language narrations, which supply temporally and logically coherent event chains, further supported by paired ego–exo views. Besides, all categories are carefully organized into 2-, 3-, and 4-event prompt settings with explicit event segmentation and ordering, with extensibility to more. This design makes MEve novel and complementary to previous benchmarks. Specifically, we construct MEve as follows:

LLM as data synthesizer Following prior settings in Kara et al. (2025), we use Gemini 2.5 Pro (Comanici et al., 2025) as a controlled prompt generator and author 60 event narratives (e.g., "$\texttt{event}_1$ then $\texttt{event}_2$ then …"). The prompts target the *General* domain and are written to avoid specific scene priors or content biases.

Prompts of diagnostic content To inherit general evaluation while disentangle how subject and action factors shape multi-event generation, we adapt VBench 2.0 categories (Zheng et al., 2025a; Zhang et al., 2024) that are critical in multi-event settings. Specifically, we convert prompts from categories *Motion Order Understanding*, *Human Identity*, and *Complex Plot* into four sequential events per prompt while preserving each category's diagnostic intent.

Prompts of real-world reference To maintain real-world temporal and logical coherence between events and isolate the effect of *viewpoint* on event controllability, we construct paired egocentric and exocentric variants of prompts from video narrations from Ego–Exo4D (Grauman et al., 2024). For viewpoint control, each prompt is prepended with one of two textual formats: (a) *Egocentric:* Add the prefix "Generate a first-person view video...", where the original subject *C* is rephrased as "the

camera wearer". (b) *Third-person:* Add the prefix "Generate a third-person view video...", denoting an external viewpoint, where the original subject C is transformed into "the person". This process results in 50 pairs of prompts, yielding in total 100 prompts.

B Implementation Details

In this work, we apply TunerDiT on the representative large open-source models Open-Sora 1.2 (Zheng et al., 2024), Open-Sora 2.0 (Zheng et al., 2025b), Wan2.1 (Team Wan et al., 2025), and Wan2.2 (Team Wan et al., 2025), spanning across different DiT structures, e.g., unified attention and factorized attention. Multi-event generation is evaluated on the proposed MEve dataset in a training-free setting. To ensure a fair evaluation, each model generates videos using its native supported frame count and resolution specifications when provided with identical text prompts.

Under the restriction of the same GPU resource, the number of frames of videos in different event counts and the resolution of the generated videos are shown in Table 2.

Table 2: Number of frames of different event count and frame resolution

Model Name	**#2**	**#3**	**#4**	**Resolution**
Open-Sora 1.2	173	201	205	1280×720
Open-Sora 2.0	204	204	204	256×256
Wan2.1	81	81	81	832×480
Wan2.2	97	97	97	1280×720
MEVG	32	48	64	256×256
DiTCtrl	97	145	193	720× 480
FreeNoise	192	192	192	256× 256

Event Partitioned Mask Open-Sora 1.2 employs ST-DiT with separate temporal and spatial blocks for self-attention and a shared cross-attention layer between text embedding and video embedding. Open-Sora 2.0 utilizes a dual-stream DiT architecture that concatenates text and video embeddings before a unified self-attention. This self-attention on the new embedding incorporates cross-attention, video self-attention, and text self-attention simultaneously. For ST-DiT in Open-Sora 1.2, the Event Partitioned Mask is applied to the temporal blocks. For dual-stream DiT in Open-Sora 2.0, the Event Partitioned Diagonal Mask operates on both blocks.

Gates and onset steps. PF conditions every segment on the first-event prompt for the first τ_{PF} denoising steps, and EM is disabled for the first τ_{EM} steps and enabled for the rest. The turning point is a relative position in the denoising trajectory that differs across DiTs, so τ_{PF}, τ_{EM}, and the band ratio r are per-model hyperparameters searched in a window around $\tau^\star/N$ and listed in Table 3.

Table 3: Per-backbone gates and band ratio of TunerDiT. Gates are fractions of the N denoising steps, i.e., $\tau = \lfloor xN \rfloor$.

Backbone	τ_{PF}/N	τ_{EM}/N	r
Open-Sora 1.2	0.1	0.1	0.1
Open-Sora 2.0	0.1	0.2	0.2
Wan2.1	0.1	0.3	0.142
Wan2.2	0.1	0.2	0.2

Text embedding partitioning. Event prompts are embedded separately and concatenated together to enable proper event segmentation and ordering through the Event Partitioned Mask. To maintain consistent positional embeddings, text embeddings are zero-padded to the maximum sequence length.

Inter-event bandwidth. Consider a prompt composed of E distinct events. Given a video with F_l latent frames in total, a video segment of F_l/E latent frames is allocated to each event. With band ratio r, the inter-event transition bandwidth is $H_{\text{inter}} = W_{\text{inter}} = \frac{rF_l}{E}$, with r per backbone in Table 3.

Cross Event Prompt Fusion All four backbones use $\tau_{\text{PF}} = \lfloor 0.1N \rfloor$, where N is the total number of denoising steps (Table 3).

C Human Evaluation

Evaluation Details Following human evaluation settings in (Cai et al., 2025), to ensure a comprehensive evaluation, we selected samples with three distinct event counts of 2, 3, and 4 from six different datasets, for a total of 18 prompts. For each prompt, we choose one sample across 8 models, yielding a total of $8 \times 3 \times 6 = 144$ videos for human assessment. Figure 8 shows our interface. All models are anonymized and randomly shuffled. Participants rate each video on the following four metrics on a 5-point Likert scale, where 1 indicates the lowest satisfaction and 5 the highest:

1. Overall Preference: *Please rate your overall preference for the video.*
2. Motion Naturalness: *How natural and realistic is the motion in the video?*
3. Transition Smoothness: *How smoothly does the video content transition across different frames?*
4. Text-Video Alignment: *To what extent does the video content match the given text descriptions?*

Evaluation Results Table 1b reports the ratings of 14 participants holding at least a bachelor's degree. TunerDiT achieves superior Text-Video Alignment scores compared to the baseline models, including the training-free methods MEVG, DiTCtrl, FreeNoise, and SwitchCraft, and TunerDiT (Wan2.1) obtains the highest overall preference, motion naturalness, and text–video alignment ratings. Furthermore, TunerDiT exhibits strong improvement over its base models with high gains across all questions, validating its acceptance by human evaluators.

D Discussion on Metrics

TA and TIS: Frame-wise and Shot-wise Base models that fuse events by default, blending backgrounds and subjects from multiple events across the entire video, can obtain spuriously high text alignment (TA) and text–image similarity (TIS) scores, because every frame then reflects all events at once. Therefore, it is necessary to emphasize both metrics: TA, which is computed with ViCLIP on a shot-wise alignment, and TIS, which is computed with frame-wise alignment. This explains our larger gain on TA.

BC, IC, and MS: Consistency Metric Hacking These default settings that models could not distinguish events in generation influence the BC, IC, and MS scores severely. Fusing every cue in the prompts on every frame, regardless of the temporal event order, will lead to the highest consistency score, while breaking generation into event segments inevitably lowers it, although more semantically correct multi-event videos are generated and aligned to human evaluation. We define this phenomenon as *Consistency Hacking*, that blending, or generally static backgrounds and subjects, increases MS due to minimal inter-frame change as well as BC and IC due to unchanged appearance, even though the video fails at multi-event separation.

Metric hacking is further exacerbated by methods that can separate events without deliberately noticing this. For example, Mask2DiT (Qi et al., 2025), although generating events separately, generates backgrounds and subjects that remain largely static in each event. This further indicates that high BC, IC, and MS scores alone do not indicate a correct multi-event video.

Trade-off between Alignment and Consistency Given the ablation study in Figures 6a and 6c, the incorporation of Prompt Fusion facilitates smoother motion dynamics during event transitions within the video. However, it does not necessarily improve TA scores, since frames sampled from transition segments between consecutive events align semantically less to both the preceding and the subsequent event, thereby it trades alignment with consistency.

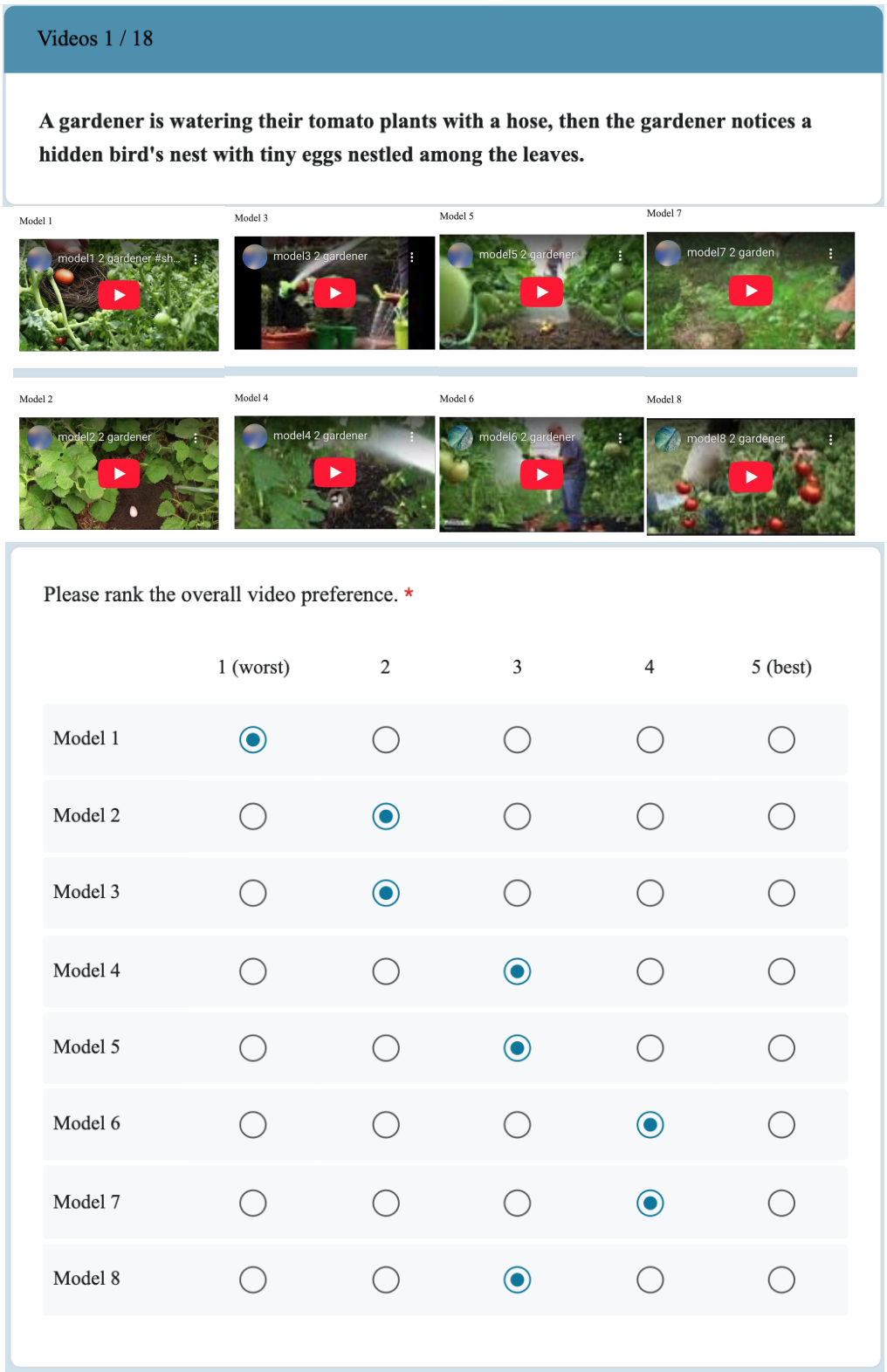


Figure 8: Interface of annotating multi-event generation quality.

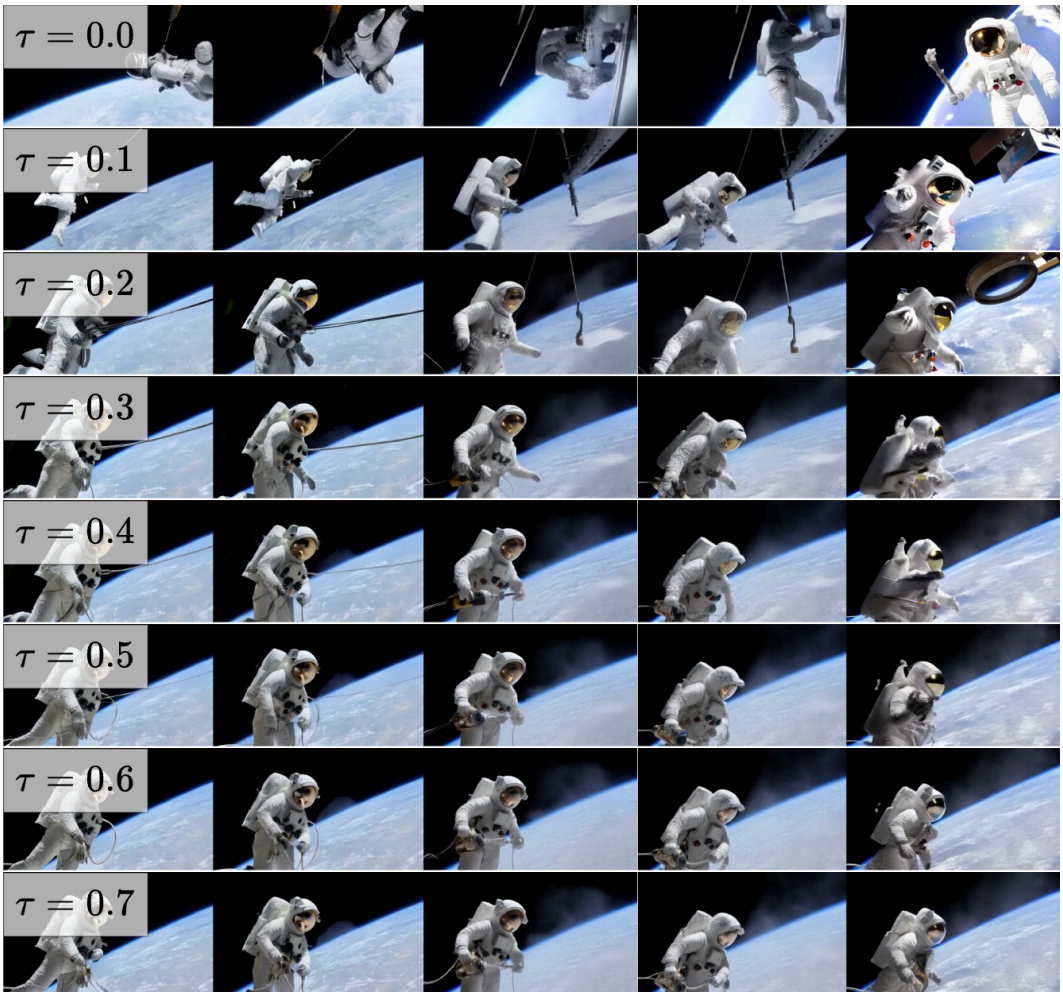


Figure 9: Generated frames as the EM gate τ_{EM} grows from 0 to $0.7N$.

E More Qualitative Analysis

As shown in Figure 9, we visualize the generation under EM gates $\tau_{\text{EM}}/N \in \{0, 0.1, 0.2, \ldots, 0.7\}$. Events begin to merge when $\tau_{\text{EM}} > 0.3N$ in the generated videos, reflecting the plateau effect in Figure 3c. This observation empirically confirms the finding that influence transitions of text conditions occur at approximately 30% of the denoising steps in DiT models. When EM starts too late ($\tau_{\text{EM}} > 0.3N$) to provide enough event segmentation, multiple events merge indistinguishably, and gradually deteriorate the performance back to the original base model.

F Additional Discussions

Investigation on Dataset Categories We compare zero-shot methods with our *TunerDiT* on Open-Sora 1.2 and Open-Sora 2.0 as shown in Fig. 10. TunerDiT ranks first on nearly all bars in BC and IC and shows consistent gains in TIS. IC and BC see the largest gains. Event-partitioned masking preserves per-event identity while prompt fusion stabilizes late-stage detail, yielding higher person and background consistency. *For motion order (MO)*, TIS increases, but BC/IC gains are moderate, which is consistent with the fact that MO stresses temporal ordering more than appearance persistence. *From viewpoint control* Generation from both views benefits from TunerDiT. 1st-person videos show marked IC improvements due to consistent hands and objects, while 3rd-view shows strong BC due to more stable backgrounds.

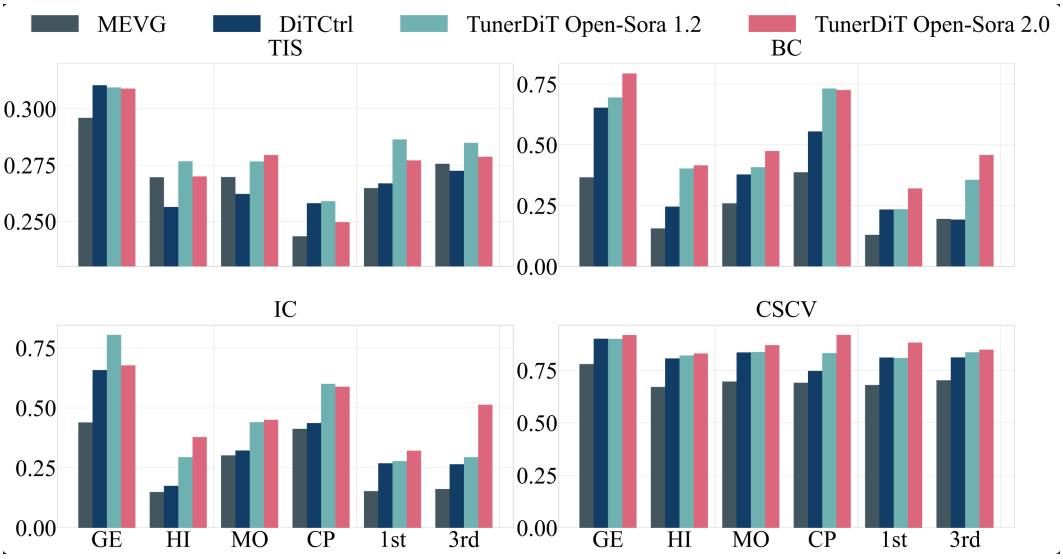


Figure 10: Investigation on dataset categories. Notation: GE (general prompts), HI (human identity), MO (motion order), CP (complex plots), and view splits (1st, 3rd).

Implications and guidance. (1) The largest gains occur in HI and CP, where boundary control and transition bands are most critical; (2) For MO-heavy prompts, later-onset prompt fusion is effective, while mask ratio can be reduced to avoid over-isolation; (3) Every TunerDiT variant exceeds every training-free baseline in MS at every event count (Table 1a), indicating that local cross-event bands suffice for smooth transitions even under camera changes.

Failure Cases We identify two primary failure modes in TunerDiT (Figure 11): *excessive fusion* and *insufficient fusion*. The former yields results resembling the original baseline models, where multiple events become indistinguishable or erroneously blended into a single sequence. Conversely, the latter leads to semantic discontinuities, such as inconsistent subjects or backgrounds across event boundaries, producing abrupt changes similar to those observed in the Concat baseline.

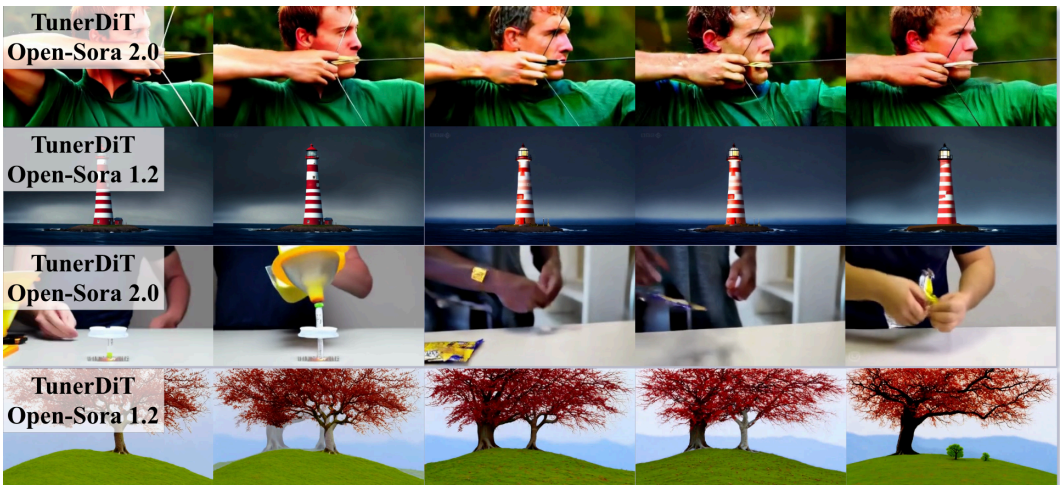


Figure 11: Visualization of failure modes in TunerDiT caused by extreme event fusion settings. The upper two rows demonstrate the effect of ***excessive fusion***, where distinct events collapse into a single, static sequence. Conversely, the lower rows illustrate ***insufficient fusion***, resulting in temporal inconsistencies regarding the subject and background across event boundaries.

Table 4: VLM-judge Event Isolation (EI) per MEve category: the accuracy of the event count that the VLM infers from sampled frames against the number of events in the prompt. Best average in **bold**, second best underlined.

Name	Average	General	HI	MO	CP	First	Third
Training-free Methods							
MEVG	0.435	0.315	0.6	0.5	0.7	0.3	0.2
FreeNoise	0.436	0.67	0.44	0.44	0.44	0.375	0.25
DiTCtrl	0.375	0.15	0.5	0.2	0.3	0.5	0.6
Mask2DiT	0.108	0.05	0.1	0.0	0.2	0.2	0.1
VideoCrafter2	0.08	0.1	0.0	0.0	0.1	0.2	0.1
SwitchCraft (Wan2.1)	0.482	0.5	0.5	0.4	0.6	0.5	0.39
Ours							
TunerDiT Open-Sora 1.2	0.499	0.2	0.6	0.3	0.5	0.8	0.6
TunerDiT Open-Sora 2.0	**0.572**	0.33	0.5	0.4	0.8	0.7	0.7
TunerDiT Wan2.1	<u>0.542</u>	0.55	0.5	0.4	0.6	0.6	0.6
TunerDiT Wan2.2	0.474	0.4	0.43	0.32	0.5	0.6	0.5

Limitations and Future Works The current implementation is constrained by the maximum text embedding length of the baseline model architecture. Given a fixed text embedding window of 512 tokens currently adopted by Open-Sora 2.0, a prompt describing E events implies that each event is allocated an average of only $\frac{512}{E}$ tokens. This token number limits the method's scalability to extremely complex sequences. However, theoretically, assuming better computational situations and the rapid development of video diffusion models for long generation, this framework can be applied naturally to longer videos with more distinct events in future due to its steering method that originates from the intrinsic perspectives shared by DiT models.

G VLM Judgement

VLM-as-a-judge evaluation. Tables 4 and 5 break down Event Isolation (EI) and Text–Video Alignment (TVA) by MEve category, with the average taken over the six categories. TunerDiT variants lead both rankings: TunerDiT Open-Sora 2.0 attains the highest average EI (0.572) and TunerDiT Wan2.1 the highest average TVA (1.617), so TunerDiT separates multiple events most reliably while keeping the strongest semantic alignment to the input prompt.

On EI, TunerDiT Open-Sora 2.0, Wan2.1, and Open-Sora 1.2 rank first to third (0.572, 0.542, and 0.499), ahead of every training-free baseline, the strongest of which is SwitchCraft (0.482). On the shared Wan2.1 backbone, TunerDiT matches or exceeds SwitchCraft in every category and raises *Third* EI from 0.39 to 0.6. The gains concentrate in the Ego-Exo4D viewpoint categories: on *First*, all four variants (0.6–0.8) exceed every baseline (at most 0.5), and TunerDiT Open-Sora 2.0 leads both *Third* (0.7) and *Complex Plot* (0.8).

On TVA, TunerDiT Wan2.1 and Open-Sora 2.0 rank first and second (1.617 and 1.533). All four variants score at least 1.49, above MEVG, FreeNoise, DiTCtrl, Mask2DiT, and VideoCrafter2 (at most 1.425). On the shared Wan2.1 backbone, TunerDiT exceeds SwitchCraft in all six categories and raises the average from 1.517 to 1.617. A TunerDiT variant holds or shares the top TVA score in five of six categories, with the largest margin on *MO*, where TunerDiT Open-Sora 2.0 and Wan2.2 reach 1.8 against at most 1.5 for baselines, consistent with the event-ordered steering.

Together, EI and TVA show that TunerDiT improves both the *structural correctness* and the *semantic faithfulness* of multi-event videos across prompt categories, and the same-backbone comparison with SwitchCraft attributes these gains to the steering of TunerDiT.

Table 5: VLM-judge Text–Video Alignment (TVA) per MEVE category: the VLM captions sampled frames, and TVA scores the alignment between these captions and the original multi-event prompt. Best average in **bold**, second best underlined.

Name	Average	General	HI	MO	CP	First	Third
Training-free Methods							
Mask2DiT	1.275	1.95	1.2	1.0	1.3	1.0	1.2
DiTCtrl	1.425	2.25	1.5	1.5	1.2	1.1	1.0
MEVG	1.375	1.35	1.8	1.4	1.3	1.0	1.4
FreeNoise	1.4	1.8	1.7	1.5	1.0	1.0	1.4
VideoCrafter2	1.183	1.5	1.1	1.2	1.3	1.0	1.0
SwitchCraft (Wan2.1)	1.517	1.9	1.6	1.5	1.5	1.3	1.3
Ours							
TunerDiT Open-Sora 1.2	1.492	1.95	1.5	1.6	1.2	1.3	1.4
TunerDiT Open-Sora 2.0	<u>1.533</u>	2.1	1.8	1.8	1.0	1.4	1.1
TunerDiT Wan2.1	**1.617**	2.0	1.7	1.6	1.6	1.4	1.4
TunerDiT Wan2.2	1.503	1.72	1.6	1.8	1.2	1.2	1.5