This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SwitchCraft: Training-Free Multi-Event Video Generation with Attention Controls

Qianxun Xu[1,2] Chenxi Song[1,†] Yujun Cai[3] Chi Zhang[1,*]

[1]Westlake University [2]Duke Kunshan University [3]The University of Queensland

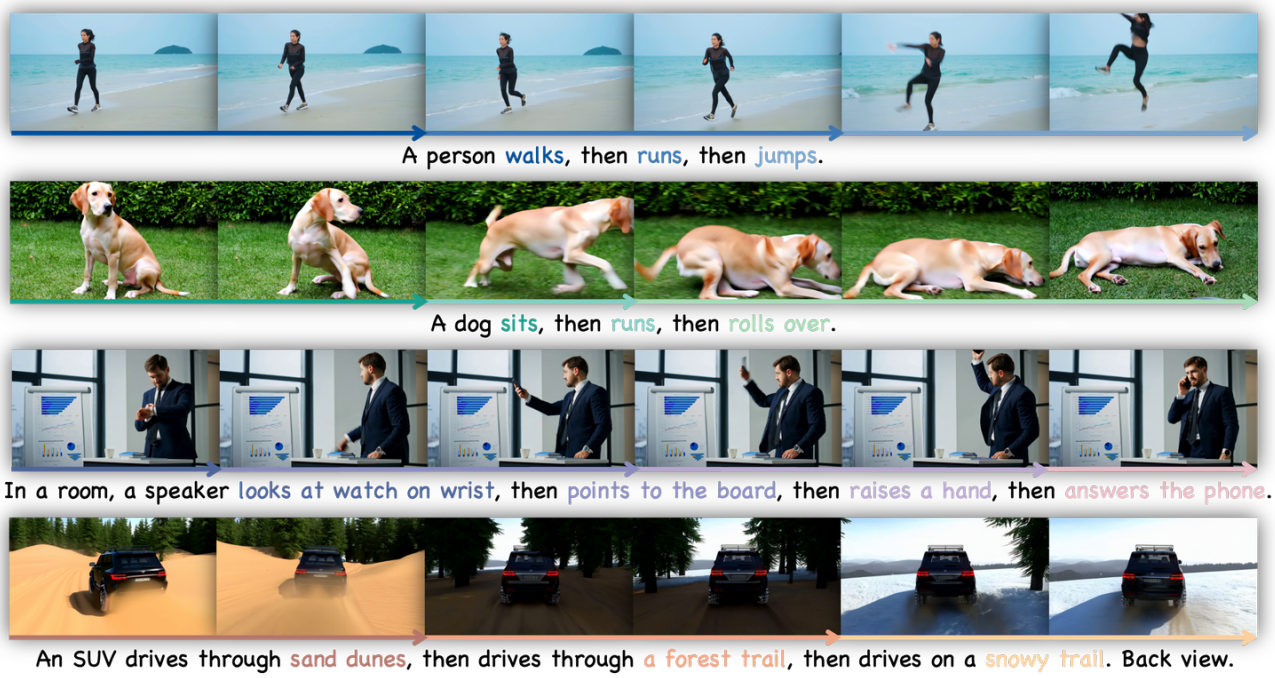

Figure 1. **SwitchCraft** enables flexible multi-event video generation across multiple actions and scenes with smooth transitions. It steers attention to enhance prompt alignment and maintain coherent temporal evolution while preserving global information.

Abstract

Recent advances in text-to-video diffusion models have enabled high-fidelity and temporally coherent video synthesis. However, current models are predominantly optimized for single-event generation. When handling multi-event prompts, without explicit temporal grounding, such models often produce blended or collapsed scenes that break the intended narrative. To address this limitation, we present SwitchCraft, a training-free framework for multi-event video generation. Our key insight is that uniform prompt injection across time ignores the correspondence between events and frames. To this end, we introduce Event-Aligned Query Steering (EAQS), which steers frame-level attention to align with relevant event prompts. Furthermore, we propose Auto-Balance Strength Solver (ABSS), which adaptively balances steering strength to preserve temporal consistency and visual fidelity. Extensive experiments demonstrate that SwitchCraft substantially improves prompt alignment, event clarity, and scene consistency compared with existing baselines, offering a simple yet effective solution for multi-event video generation. Code will be released here.

1. Introduction

Recent years have witnessed remarkable progress in text-to-video (T2V) generation powered by diffusion models.

†Project lead. *Corresponding author.

Large-scale pretrained systems can now synthesize visually compelling video clips that align well with textual descriptions, exhibiting strong visual fidelity, temporal coherence, and identity consistency across diverse scenes [14, 15, 22, 37]. These advances have significantly democratized video generation, making it central to a broad spectrum of applications [1, 8, 18, 19, 21, 23, 28, 35, 36, 38].

Despite these advances, current state-of-the-art T2V models are predominantly optimized for generating single-event videos guided by one global caption prompt. In other words, they excel when the prompt describes a single coherent scene or action. However, when the prompt specifies multiple temporally ordered events, existing models often struggle to capture the compositional and temporal structure implied by the text. Instead of generating a sequence that faithfully follows the described narrative, the resulting videos tend to exhibit events blending, ambiguous transitions, or omit events and collapse into dominant scenes. This limitation stems from how current architectures process text guidance during video generation. Typically, they apply the same caption representation uniformly across all timesteps through mechanisms such as cross-attention. As a result, the model interprets the entire description as a holistic context rather than a temporally structured sequence of events. This uniform conditioning neglects the intrinsic temporal hierarchy in complex prompts, making it difficult for the model to reason about event boundaries, causal order, or evolving scene dynamics.

Existing studies have attempted to introduce temporal control from two main directions. The first relies on training or fine-tuning the generation backbone with temporally grounded supervision, where each event is explicitly aligned with a specific time span [39, 40]. While this improves temporal correspondence between actions and frames, it requires densely annotated data, incurs heavy computational costs, and generalizes poorly to unseen or loosely structured scenarios. The second line adopts a concatenation strategy that fuses separately generated clips, leveraging information from previous segments to create smoother transitions [3, 28, 33, 34]. However, this clip-wise process lacks global context, so each segment has no foresight of subsequent events and cannot stage objects, poses, or motion cues in advance. Transitions thus become discontinuous and exhibit temporal drift across segments. Consequently, these approaches still fall short of achieving flexible multi-event video generation, where multiple actions evolve coherently within a single continuous generation process, maintaining temporal order and scene consistency without retraining the model.

Motivated by these limitations, we aim to address the challenge of multi-event video generation by optimizing how the model attends to different events over time. We identify that the core issue lies in the way prompts are injected through cross-attention: existing models lack a mechanism to accurately associate each event prompt with its corresponding temporal segment in the video. Without such alignment, attention becomes globally entangled, causing interference between unrelated events.

To overcome this, we introduce Event-Aligned Query Steering (EAQS), a training-free mechanism that aligns frame level attention with the corresponding event description. Instead of uniformly injecting the entire text prompt across all frames, EAQS modulates the query vectors of each frame using compact event-specific key subspaces built from textual anchor tokens. It increases the projection of queries onto the target event subspace and suppresses components aligned with competing events, which reduces event interference and improves disentanglement without changing model weights or training.

The steering strength is crucial for temporal control. Overly aggressive query perturbations may distort subject appearance or destabilize motion, whereas insufficient steering can fail to override the inherent bias of pretrained T2V models toward globally blended scenes. To robustly balance these competing factors, we propose the Auto-Balance Strength Solver (ABSS), an inference-time optimization module that adaptively computes enhancement and suppression coefficients for EAQS. ABSS compresses event-level directions via singular value decomposition, analyzes the current query–key alignment margins between the target event and all competing events, and solves a small convex optimization problem to obtain scale aware strengths that enlarge the target margin while limiting query displacement. This yields stable steering that preserves temporal coherence and visual fidelity with no additional hyperparameters or prompt-specific tuning.

To validate the effectiveness of our approach, we conduct extensive experiments on multiple benchmarks. The results demonstrate that our method significantly outperforms both baseline and state-of-the-art models in generating coherent, temporally ordered multi-event videos while maintaining high visual quality. Our contributions are summarized as follows:

- We propose **SwitchCraft**, a training-free framework enabling time-controlled multi-event video generation with clear event ordering and smooth transitions.
- We design **Event-Aligned Query Steering (EAQS)** to dynamically align frame-level attention with relevant event prompts for accurate temporal localization.
- We introduce **Auto-Balance Strength Solver (ABSS)** to adaptively balance steering strength, maintaining temporal and visual consistency.

2. Related Work

Text-to-video diffusion models. T2V diffusion models have advanced rapidly in visual quality, duration and con-

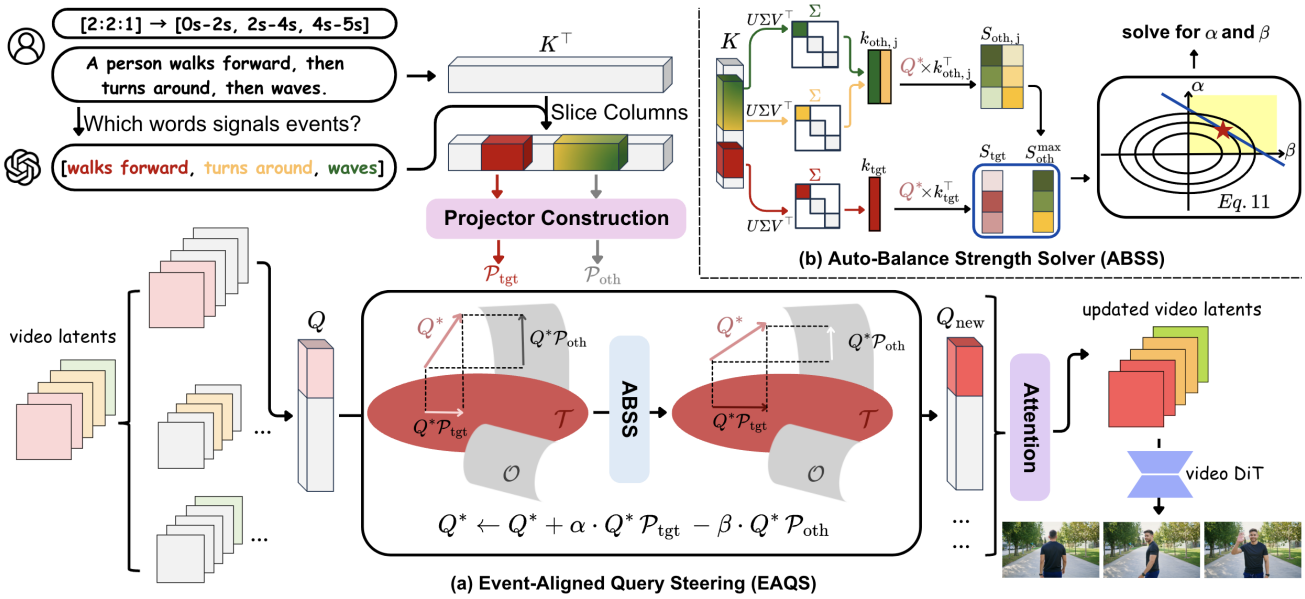

Figure 2. **Overview of SwitchCraft.** (a) **EAQS** takes a text prompt and user specified event time spans, identifies anchor tokens for each event, and constructs event specific projectors from their attention keys. It steers video queries toward the target event and away from others in each temporal span. (b) **ABSS** estimates enhancement and suppression strengths by extracting dominant directions from the event keys and correcting the attention deficit. The updated queries pass through the video diffusion transformer so that each temporal span follows its intended event with smooth transitions.

trollability [4, 5, 9, 10, 30, 36, 42, 44]. Early methods adapt pretrained image diffusion models to videos by inserting temporal attention and extending U-Net style denoisers to operate over space and time [2, 4, 12, 14, 15, 32, 36]. More recent architectures replace the U-Net backbone with transformer based diffusion models that treat video as a sequence of space time tokens, which improves scalability to higher resolution, longer duration and more complex motion [1, 27, 41].

Another line of work adopts an autoregressive formulation. Diffusion transformers are adapted into causal generators with key value caching and streaming inference, enabling low latency, frame by frame rollout for real time video generation [7, 16, 40, 43]. This improves responsiveness but also makes long term stability difficult, since small early errors can accumulate and cause drift. Recent methods address this exposure bias by training on the model's own streamed rollouts and by using rolling window objectives, cache sharing, and attention anchoring to preserve identity and scene layout over extended sequences [6, 7, 16, 24]. These advances improve duration and streaming quality, but they do not explicitly organize multiple semantically distinct events. Our method addresses this goal directly.

Multi-event video generation.

Multi-event video generation aims to produce one video with several events in a specified order while preserving scene layout, identities and motion [5, 20, 25, 45, 46]. A common strategy is sequential stitching. Phenaki [33], Gen-L-Video [34] and MEVG [28] generate one segment per prompt, conditioning each new segment on its text and on visual context from previous frames to keep appearance and background consistent. This improves continuity but offers no global view of all events and no explicit control over event duration or boundaries. LongLive [40] extends this paradigm to minute scale streaming videos with cached features, but its causal design blocks access to future events and the accumulated history induces prompt inertia and long range drift. DiTCtrl [3] instead edits attention in a pretrained diffusion transformer to enforce new prompts over time, but this often harms scene consistency and visual quality. Mind the Time [39] binds event captions to explicit time spans and fine tunes a diffusion transformer with time aware cross attention, which gives temporally grounded control but demands expensive fine tuning and dense temporal labels. In contrast, our method is training free and provides global control over multiple semantically distinct events while maintaining coherence across the entire video.

3. Method

SwitchCraft controls how a pretrained T2V diffusion model attends to multiple events over time. We build on a diffusion transformer backbone and modify its cross attention at inference to align each temporal span with its designated

event. Section 3.1 reviews the backbone and its attention mechanism. Section 3.2 introduces EAQS, which defines an event-specific query update that strengthens alignment with the text features of the target event while suppressing responses to other events. Section 3.3 then presents ABSS, which infers the corresponding enhancement and suppression strengths dynamically and supplies them to EAQS. Figure 2 provides an overview of our pipeline.

3.1. Preliminaries

Text-to-video Diffusion Transformers (DiTs). Given a video with F frames and size $H \times W$, a video tokenizer encodes it as $z \in \mathbb{R}^{C \times F' \times H' \times W'}$ with temporally and spatially downsampled dimensions. The latent tensor is reshaped into a token sequence, and a text conditioned diffusion transformer maps noisy latents z_t toward the clean latents using a velocity prediction under text conditioning y:

$$\mathcal{L} = \mathbb{E}_{z_t,t,y}\big[\| u_\theta(z_t, t, y) - v_t^\star \|_2^2\big], \tag{1}$$

where u_θ is the transformer denoiser and $v_t^\star$ is the target velocity.

Cross attention mechanism. T2V DiTs inject prompt information into video features through cross attention. Let $X \in \mathbb{R}^{S \times d}$ be the video token sequence at the current denoising step and $Y \in \mathbb{R}^{N \times d}$ the text tokens from the prompt. The layer forms queries from video and keys and values from text with trainable weight matrices W_Q, W_K, W_V,

$$Q = XW_Q\,, \quad K = YW_K\,, \quad V = YW_V\,, \tag{2}$$

then measures video–text alignment with scaled dot products and injects text guidance back into the video stream,

$$A = \mathrm{softmax}\big(d_h^{-\frac{1}{2}} QK^\top\big)\,, \quad \mathrm{out} = AV\,. \tag{3}$$

Here d_h is the per-head feature dimension and out is the cross-attention output that updates the video tokens. In current models this attention is shared across all frames, so a text token that describes a single event can affect every timestep. SwitchCraft instead edits the frame queries within each temporal span, steering attention so that each event primarily governs its designated segment.

3.2. Event-Aligned Query Steering (EAQS)

EAQS defines a span-local query update that is controlled by two non-negative steering strengths for enhancement and suppression. In this subsection we treat these strengths as given and focus on how to construct event specific directions from the text keys and how to use them to modify video queries inside the corresponding temporal spans. The procedure that selects the strengths for each span is described in Section 3.3.

EAQS enforces event specific temporal alignment in three steps: it decomposes the prompt into distinct events and extracts anchor tokens for each, assigns each event to a dedicated latent temporal window, and updates queries within each window to increase alignment with the target anchors while suppressing competing events.

Event-specific anchor identification. We use a large language model such as ChatGPT [29] to extract event anchors, defined as a small set of words or short phrases that distinguishes one event description from the others. For example, in scene transitions the anchors focus on setting descriptors such as sunny desert or icy cave, and in multi-event behavior they focus on concise action phrases such as walking forward or reading a book. The instruction prompt that we provide to the language model can be found in the supplementary material. These anchors are then mapped through the backbone tokenizer and recorded as index sets of text tokens for that event.

Temporal window assignment. After extracting event phrases, SwitchCraft assigns each event to a temporal span in the latent space. The user may specify relative duration weights $\{w_1, \ldots, w_A\}$ for the A events, or choose to default to an equal split. For a video of F' latent frames, the discrete frame count N_i allocated to the i-th event is proportionally defined as:

$$N_i \approx F' \cdot \frac{w_i}{\sum_{j=1}^{A} w_j}. \tag{4}$$

To ensure exactly F' frames are assigned without gaps, we round these targets to integers and distribute any truncation remainder to the events with the largest fractional parts. The i-th event is then mapped to the contiguous, half-open frame index interval $[\sum_{j=1}^{i-1} N_j, \sum_{j=1}^{i} N_j)$.

Query Steering. As defined in Equation (3), cross attention governs how each frame draws information from the text tokens. To strengthen alignment with the text, we aim to increase the values of $QK^\top$ associated with the target tokens within the event span and to reduce the entries associated with tokens from other events. Since keys and values are shared across frames, modifying them would propagate effects to all frames and can degrade frames not intended to change. We therefore edit only the queries, which are frame-specific and can steer attention only within their assigned event span.

We steer queries at the level of individual heads to obtain fine grained temporal control. Inside one cross attention block and for a single head, let $K \in \mathbb{R}^{L_k \times D}$ be the matrix of text keys and let $Q^* \in \mathbb{R}^{R \times D}$ be the slice of query rows for the latent frames assigned to the current event, with L_k the number of text tokens, D the feature dimension, and R the number of rows corresponding to the video tokens in this span. We form K_{tgt} by gathering from K the rows corresponding to the target event's anchor tokens, and form

K_{oth} by gathering the rows corresponding to anchor tokens from the remaining events. We want every row of Q^* to produce higher dot products with columns from its corresponding $K_{\text{tgt}}^\top$ and lower dot products with columns from $K_{\text{oth}}^\top$. The most direct way to achieve this is to decompose queries into components relative to the spans of K_{tgt} and K_{oth}, then enhance the component that lies in the span of K_{tgt} and suppress the component that lies in the span of K_{oth}.

To carry out this decomposition we need an operator that returns, for any query, its component within a specified key span. A naive attempt is to compute $Q^* \cdot K^\top$, but this reports token-level instead of event-level alignment, and lives in $\mathbb{R}^{R \times L_k}$ which is not the same space as $Q^* \in \mathbb{R}^{R \times D}$. We therefore construct right projectors P_{tgt} and P_{oth} from K_{tgt} and K_{oth}, that can project Q^* to their key spans defined by

$$P_{\text{tgt}} \;=\; \mathcal{P}[K_{\text{tgt}}] \;=\; K_{\text{tgt}}^\top \big(K_{\text{tgt}} K_{\text{tgt}}^\top + \epsilon I\big)^{-1} K_{\text{tgt}}, \qquad (5)$$

with small ridge $\epsilon > 0$ and identity I, and analogously $P_{\text{oth}} = \mathcal{P}[K_{\text{oth}}]$. We view these spans as two subspaces in the D-dimensional key space: $\mathcal{T}$ for the target event, spanned by K_{tgt}, and $\mathcal{O}$ for all competing events, spanned by K_{oth} (see Figure 2). The projectors P_{tgt} and P_{oth} therefore map Q^* onto $\mathcal{T}$ and $\mathcal{O}$, respectively.

To reinforce the target event within its temporal span and reduce the influence of unrelated events, we modulate the queries with nonnegative strengths α and β:

$$Q^* \leftarrow Q^* + \alpha \cdot Q^* P_{\text{tgt}} - \beta \cdot Q^* P_{\text{oth}}. \qquad (6)$$

The first term boosts the component of Q^* that lies in the span of $\mathcal{T}$, which increases alignment with the keys of the intended event. The second term subtracts the component of Q^* that lies in the span of $\mathcal{O}$, which weakens alignment with keys that correspond to other events and reduces leakage. After this edit we apply a row-wise renormalization to stabilize attention magnitudes. This update shifts the queries of frames in the span toward the target event, so the generated content follows the intended instruction while the influence of unrelated events is suppressed.

Notably, we apply the edit only in the early denoising steps and in earlier blocks. This choice is motivated by how diffusion transformers organize generation over time and depth. Early denoising steps and early transformer blocks establish scene layout and large scale motion, while later steps and later blocks primarily refine texture, identity, and appearance detail [26, 27, 41]. Steering in this early stage is therefore sufficient to enforce when each event should occur. After this point the unmodified base model can complete high frequency detail without further intervention. Additionally, because the update acts on the queries rather than on the attention weights after the softmax, it preserves the structure learned by the pretrained model and avoids abrupt changes at event boundaries.

3.3. Auto-Balance Strength Solver (ABSS)

To adapt the steering strength to different prompts and scenes, we introduce ABSS, which robustly determines the enhancement α and suppression β used in the update in Equation 6. The goal is to ensure that the current event dominates attention within its assigned temporal span while avoiding excessive edits that would distort appearance or destabilize motion. ABSS operates at inference and measures, on the frames of the active span, the relative alignment of the current queries with the target span K_{tgt} against the competing span of K_{oth}. It then solves a two-variable constrained quadratic program to select nonnegative α and β.

We aim to correct the deficit in each event span. A direct comparison based on $\Delta = Q^* K_{\text{oth}}^\top - Q^* K_{\text{tgt}}^\top$ and optimizing from Δ would be high dimensional and expensive to evaluate, sensitive to token count and order, and does not yield a stable feature space edit. We therefore compress the keys of each event into one dominant direction by applying singular value decomposition to the normalized key rows. Denote these unit directions by k_{tgt} for the current event and $k_{\text{oth},j}$ for each competitor event $j = 1, \ldots, J$. This yields one representative direction per event that is robust to the number and location of anchor tokens. Let $k_{\text{oth}} = [\, k_{\text{oth},1} \;\cdots\; k_{\text{oth},J} \,] \in \mathbb{R}^{D \times J}$ denote the matrix whose columns are the competitor directions $k_{\text{oth},j}$. We then define row-wise alignment scores $S_{\text{tgt}} \in \mathbb{R}^R$ and $S_{\text{oth}} \in \mathbb{R}^{R \times J}$ for $Q^* \in \mathbb{R}^{R \times D}$ by

$$S_{\text{tgt}} = Q^* \, k_{\text{tgt}}, \quad S_{\text{oth}} = Q^* \, k_{\text{oth}} \,. \qquad (7)$$

For every row, take the dominant competitor $S_{\text{oth}}^{\max} = \max_j S_{\text{oth},j}$ and define the margin deficit

$$d = S_{\text{oth}}^{\max} - S_{\text{tgt}} + \varepsilon \,. \qquad (8)$$

To solve for the steering strengths α and β we formulate a convex minimization problem that balances two requirements. The first is to make the target event dominate within the current span by raising the margin against competing events. The second is to prevent oversteering by discouraging unnecessarily large values. For the first requirement, to compensate for the deficit d we require that the combined effect of enhancement and suppression satisfies:

$$\alpha \, S_{\text{tgt}} \;+\; \beta \, S_{\text{oth}}^{\max} \;\geq\; d. \qquad (9)$$

Additionally, to keep the edit conservative and protect appearance, we introduce a resistance matrix that measures how expensive it is to move along each steering direction. We penalize large strengths using a diagonal matrix M whose entries reflect how sensitive the margin is to changes in α and β over the span:

$$M = \begin{bmatrix} \|S_{\text{tgt}}\|_2^2 & 0 \\ 0 & \|S_{\text{oth}}^{\max}\|_2^2 \end{bmatrix}. \qquad (10)$$

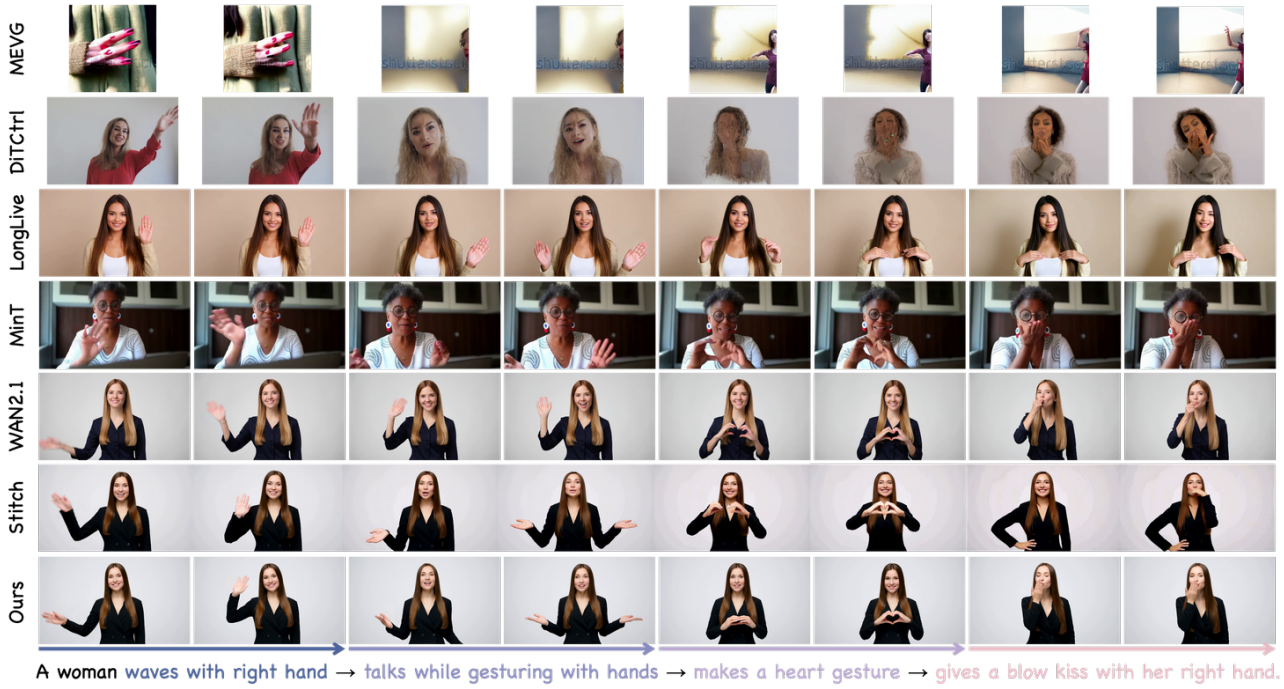

Figure 3. **Qualitative comparison.** SwitchCraft executes all events in the intended order, prevents leakage and omission, and maintains subject and scene consistency. Baselines show omissions, drift across spans, or progressive quality decay.

We penalize large strengths using a quadratic term $x^\top M x$ with $x = [\alpha \quad \beta]^\top$. The diagonal entries of M quantify how strongly changes in α and β affect the margin over the span, so directions that already induce large margin changes receive a higher cost. This yields scale-aware damping, stabilizes the update, and avoids pushing too far.

Finally, with $C = [\, S_{\text{tgt}} \;\; S_{\text{oth}}^{\max}]$, the convex minimization problem is formulated by:

$$\min_{x \geq 0} \; \frac{1}{2} x^\top M x \; + \; \frac{1}{2} \left\| \max\!\big(0,\, d - Cx\big) \right\|_2^2. \tag{11}$$

The second term is inactive when $d - Cx \leq 0$, in which case the minimum is $x = 0$. When it is active the unique stationary point satisfies

$$\big(M + C^\top C\big)\, x \;=\; C^\top d, \quad x \leftarrow \max(x, 0). \tag{12}$$

The full derivation can be found in the supplementary material. The final α and β are applied uniformly within the current event span.

4. Experiments

4.1. Experimental Setup

We implement SwitchCraft on the Wan 2.1 T2V 14B backbone [37] and generate videos with a standard resolution of 832×480 pixels and length of 5 seconds on one NVIDIA A100 GPU. Sampling uses 50 denoising steps and the query steering is enabled only in the first 20 blocks of the first 20 steps to shape layout and motion, and edits are disabled for the remaining blocks and steps so the base model refines appearance. All other sampling parameters follow the Wan 2.1 default configuration.

We evaluate on 60 multi-event prompts that cover action switches and scene transitions. Prompts are drawn from prior work and from an automatically generated set produced by a large language model [29], spanning two to four events per video. We compare SwitchCraft to baselines of training-based temporal controllers MinT [39] and LongLive [40], training-free multi-event controller MEVG [28] and DiTCtrl [3], and the Wan 2.1 base model [37]. We also design a stitching baseline *Stitch* that generates consecutive segments with the Wan 2.1 image-to-video (I2V) model using the last frame of the previous segment as the starting frame. Note that MinT [39] does not release its base model, trained weights, or training data, so we compare to its public samples for qualitative analysis and human evaluation only. For all baselines, we follow their default experimental setups.

We evaluate text alignment, temporal smoothness, and visual quality and consistency. Text alignment is measured by per-frame CLIP-T [13, 31] and by the text-to-video alignment score in VideoScore2 [11]. Temporal smoothness is measured by CLIP-F, which computes cosine similarity

Table 1. **Quantitative comparison and ablation study.** SwitchCraft improves text alignment and preserves temporal smoothness and visual quality across CLIP [13, 31], VideoScore2 [11], and VBench [17]. Best scores are in **bold** and second best are underlined. Columns highlighted in yellow correspond to text alignment metrics, purple to temporal smoothness metrics, and blue to visual quality and consistency metrics. Removing components in the ablations generally reduces text alignment, while the other metrics remain close to the Wan 2.1 [37] backbone.

	CLIP [13, 31]		VideoScore2 [11]			VBench [17]				
Method	CLIP-T ↑	CLIP-F ↑	Visual quality ↑	T2V alignment ↑	Phys. cons. ↑	Motion smoothness ↑	Sub cons. ↑	Back cons. ↑	Aesthetic quality ↑	Imaging quality ↑
MEVG [28]	0.244	0.915	2.13	2.33	1.73	0.953	0.701	0.841	0.346	0.525
DiTCtrl [3]	0.246	0.959	3.20	3.27	2.93	0.981	0.764	0.876	0.511	0.702
LongLive [40]	0.252	**0.984**	4.27	3.13	3.97	0.984	0.898	0.908	0.627	0.725
Wan2.1 [37]	0.256	0.980	4.30	3.47	4.12	0.987	**0.947**	**0.924**	0.645	0.738
Stitch	0.257	0.963	3.73	3.67	3.80	0.983	0.926	0.910	0.608	0.711
Ours	**0.275**	0.980	**4.33**	**4.30**	**4.13**	**0.989**	0.945	0.921	**0.648**	**0.741**
Random strength	0.253	0.974	4.15	3.62	3.98	0.987	0.939	0.915	0.637	0.734
Fixed strength	0.264	0.967	3.97	3.75	3.95	0.985	0.934	0.907	0.631	0.715
w/o SVD	0.255	0.978	4.30	3.67	4.08	0.988	0.943	0.918	0.643	0.734
w/o enhance	0.262	0.980	4.35	3.78	4.13	0.989	0.945	0.922	0.646	0.739
w/o suppress	0.261	0.978	4.28	3.73	4.05	0.986	0.942	0.920	0.642	0.736

between CLIP image embeddings of adjacent frames [13, 31], and by the motion smoothness index in VBench [17]. Visual quality is assessed using the visual quality and physical consistency scores in VideoScore2 [11], and the subject consistency, background consistency, aesthetic quality (perceived visual appeal), and imaging quality (technical fidelity) scores in VBench [17].

4.2. Qualitative Results

Qualitative comparison. Figure 3 compares SwitchCraft with strong baselines on a four-event prompt from the released work of MinT [39]. MEVG [28] exhibits low visual fidelity due to its earlier backbone and converts each event into a new viewpoint, breaking identity and background continuity. DiTCtrl [3] shows a similar structural weakness: transitions are visually smooth, yet the global scene and subject drift from event to event. LongLive [40] produces only a small subset of the required events, actions stagnate and frame quality degrades over time. MinT [39] generates coherent actions aligned with the text but leaks activity across spans, as the subject continues talking outside the designated segment. The base Wan 2.1 [37] preserves appearance but misses later events and blurs boundaries, with the first action persisting into subsequent segments. Stitch baseline yields seemingly reasonable sequences but is highly sensitive to the previous key frame and inherits its motion bias, causing action bleed and occasional discontinuities, as corroborated by the failure in Figure 4. In contrast, SwitchCraft maintains subject and scene, executes all events in the correct order, avoids leakage and omission, and sustains visual quality over time. More qualitative results can be found in the supplementary material.

More applications. As shown in Figure 5, SwitchCraft can produce creative transition effects when the middle segment of the script describes a visual occluder that gradually masks the current scene and reveals the next one. Because the edit acts only on video queries to steer to specific events, subject identity and global context can propagate across segments while actions from adjacent segments do not bleed into the transition.

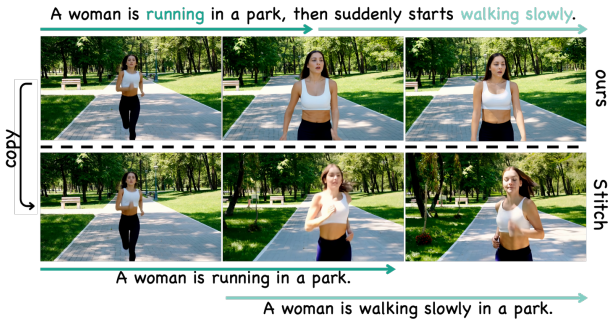

Figure 4. **Failure of *Stitch*.** Segment-wise generation inherits bias from the last frame of the previous segment and propagates motion and layout, causing action bleed and unstable handoffs.

Figure 5. **More applications.** SwitchCraft generates creative occluding transitions between scenes, preventing event bleeding while preserving identity and global context.

4.3. Quantitative Results

Quantitative comparison. We conduct a quantitative comparison between SwitchCraft and other state-of-the-

Table 2. **Human evaluation.** Users scored no omission, no leakage, transition smoothness, and visual quality on a five point scale. SwitchCraft attains the best overall performance.

Method	No omission ↑	No leakage ↑	Smooth transition ↑	Visual quality ↑
MEVG [28]	1.41	1.38	1.38	1.28
DiTCtrl [3]	1.66	1.48	1.48	1.59
LongLive [40]	2.07	2.72	2.97	3.52
MinT [39]	**4.31**	3.69	3.76	3.83
Wan 2.1 [37]	3.17	3.38	3.79	3.93
Stitch	2.62	2.07	2.14	2.45
Ours	4.21	**4.04**	**3.93**	**4.24**

art methods. Our experimental results in Table 1 demonstrate the superior performance of our method that significantly enhances text alignment without degrading motion smoothness and visual quality. Notably, our method does not achieve the highest score in CLIP-F because this metric rewards very similar neighboring frames, so larger pose changes or fast actions reduce the score even when transitions are smooth and realistic. Our scores in subject and background consistency are slightly lower than that of Wan2.1 base model [37] because VBench [17] measures these consistencies by feature similarity across frames. SwitchCraft emphasizes the requested events and reveals new viewpoints and regions, which may lower feature similarity even when identity and scene remain stable to human observers.

Human evaluation. To complement automatic metrics and assess human preferences, we conduct a user study across SwitchCraft and all baselines. We invited 29 users and use a five point scale where 5 is best. The participants are asked to score four aspects: no omission meaning all requested events appear, no leakage meaning events do not appear outside their spans, smooth event transition, and visual quality. As shown in Table 2, SwitchCraft achieves the highest mean scores on no leakage, transition smoothness, and visual quality, and is close to the training-based method MinT [39] on no omission. These results show that our method delivers the intended event sequence with fewer side effects and high perceived quality.

4.4. Ablation Study

We conduct ablations to assess the contribution of each component of SwitchCraft. As shown in Table 1 and Figure 6, the full model executes all events in the correct spans while maintaining high visual quality and consistency.

Effectiveness of ABSS. Removing ABSS and sampling random strengths from 0 to 1 causes missed or delayed events, and fixing the strength to 1 leads to over-steering and degraded appearance. Removing the SVD weakens the separation between events in key space, which lowers text alignment while keeping frame quality close to the backbone.

Effectiveness of EAQS. In EAQS, the suppression-only variant (no enhancement) reduces leakage of events but does not actively drive queries toward the next event, so later events often reduce and disappear. Enhancement-only (no suppression) fails to isolate spans when competitors are strong, so dominant actions persist and blend with the target event, introducing artifacts and reducing both alignment and perceived quality. These results show that all key components are necessary to realize the intended multi-event sequence while preserving visual quality.

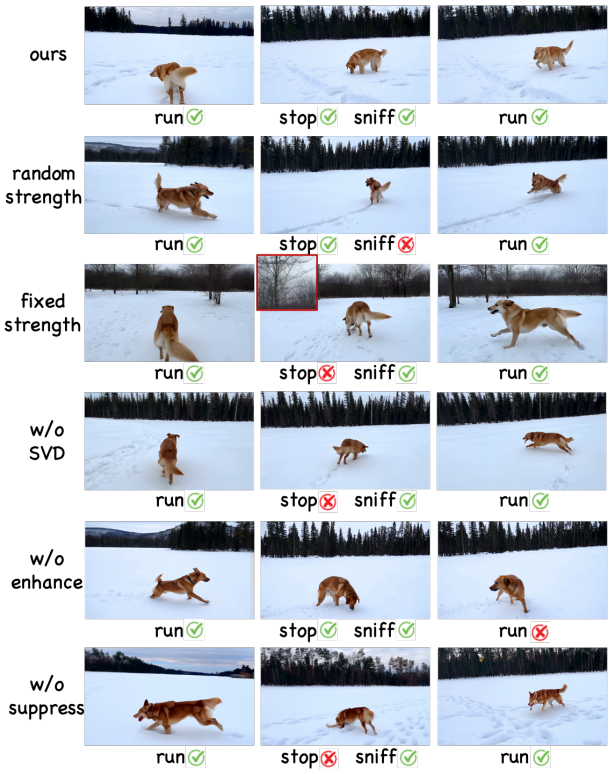

Figure 6. **Ablation study of key components.** Prompt: *On a snowy plain, a dog is running forward, then suddenly stops to sniff the ground, then continues running.* The full method yields complete sequences with stable quality. Removing parts leads to event leakage, omission, and degraded visual quality.

5. Conclusion

In this work, we presented SwitchCraft, a training-free framework for multi-event video generation. Our Event-Aligned Query Steering modifies video queries within each temporal span so that they preferentially attend to the target event and are repelled from competing events, while the Auto-Balance Strength Solver dynamically selects enhancement and suppression strengths, avoiding under- or over-steering. Experiments across automatic metrics and human evaluations show that SwitchCraft improves text alignment, temporal smoothness, and visual quality over baselines.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 6250070674) and the Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (2024R01007).

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. *ACM TOG*, 43(6), 2024. 2, 3

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 3

[3] 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 *CVPR*, 2025. 2, 3, 6, 7, 8, 4

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3

[5] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2024. 3

[6] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025. 3

[7] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-VDM: Efficient autoregressive video diffusion model with causal generation and cache sharing. In *ICML*, pages 18550–18565, 2025. 3

[8] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. In *ECCV*, 2024. 2

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[10] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *NeurIPS*, 2022. 3

[11] Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025. 6, 7

[12] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 3

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 6, 7

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022. 2, 3

[16] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 3

[17] 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 *CVPR*, 2024. 7, 8

[18] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. VMC: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *CVPR*, 2024. 2

[19] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, 2024. 2

[20] 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 *CVPR*, 2025. 3

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV*, 2023. 2

[22] Weijie Kong et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[23] Taegyeong Lee, Soyeong Kwon, and Taehwan Kim. Grid diffusion models for text-to-video generation. In *CVPR*, 2024. 2

[24] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025. 3

[25] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *ECCV*, 2024. 3

[26] Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K. Wong, Yu Qiao, and Ziwei Liu. Dual-expert consistency model for efficient and high-quality video generation. In *ICCV*, 2025. 5

[27] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Trans. Mach. Learn. Res.*, 2025. 3, 5

[28] 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 *ECCV*, 2024. 2, 3, 6, 7, 8, 5

[29] OpenAI, Joshua Achiam, Steven Adler, Sandhini Agarwal, et al. GPT-4 technical report, 2023. 4, 6

[30] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *ICLR*, 2024. 3

[31] 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. *arXiv preprint arXiv:2103.00020*, 2021. 6, 7

[32] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 3

[33] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023. 2, 3

[34] 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. 2, 3

[35] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In *CVPR*, 2024. 2

[36] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, 2025. 2, 3

[37] WanTeam, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 6, 7, 8, 1, 3, 4

[38] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 2

[39] 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 *CVPR*, 2025. 2, 3, 6, 7, 8

[40] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025. 2, 3, 6, 7, 8, 4

[41] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 3, 5

[42] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Gong Ming, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over diffusion for extremely long video generation. In *ACL*, 2023. 3

[43] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. 3

[44] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 3

[45] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *NeurIPS*, 2024. 3

[46] Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *ACM MM*, 2023. 3