TokenDial: Continuous Attribute Control in Text-to-Video via Spatiotemporal Token Offsets

Zhixuan Liu[1,2,*] Peter Schaldenbrand[2] Yijun Li[1] Long Mai[1]
Aniruddha Mahapatra[1] Cusuh Ham[1] Jean Oh[2] Jui-Hsien Wang[1]
[1]Adobe Research [2]Carnegie Mellon University

https://tokendial.github.io

Figure 1. ***TokenDial* enables continuous slider control of both appearance and motion dynamics in text-to-video generation.** Increasing slider strength produces progressive, monotonic changes in appearance (top) and motion magnitude (bottom) while preserving identity, background, and temporal coherence. Left-side labels are for illustration only (not prompts). Red circles highlight the same region to compare motion at matched time steps.

Abstract

We present TokenDial, a framework for continuous, slider-style attribute control in pretrained text-to-video generation models. While modern generators produce strong holistic videos, they offer limited control over how much an attribute changes (e.g., effect intensity or motion magnitude) without drifting identity, background, or temporal coherence. TokenDial is built on the observation: additive offsets in the intermediate spatiotemporal visual patch-token space form a semantic control direction, where adjusting the offset magnitude yields coherent, predictable edits for both appearance and motion dynamics. We learn attribute-specific token offsets without retraining the backbone, using pretrained understanding signals: semantic direction matching for appearance and motion-magnitude scaling for motion. We demonstrate TokenDial's effectiveness on diverse attributes and prompts, achieving stronger controllability and higher-quality edits than state-of-the-art baselines, supported by extensive quantitative evaluation and human studies.

1. Introduction

Text-to-video (T2V) generation has advanced rapidly, producing high-quality videos from high-level prompts. Yet for

[*]Work done while Zhixuan was an intern at Adobe Research.

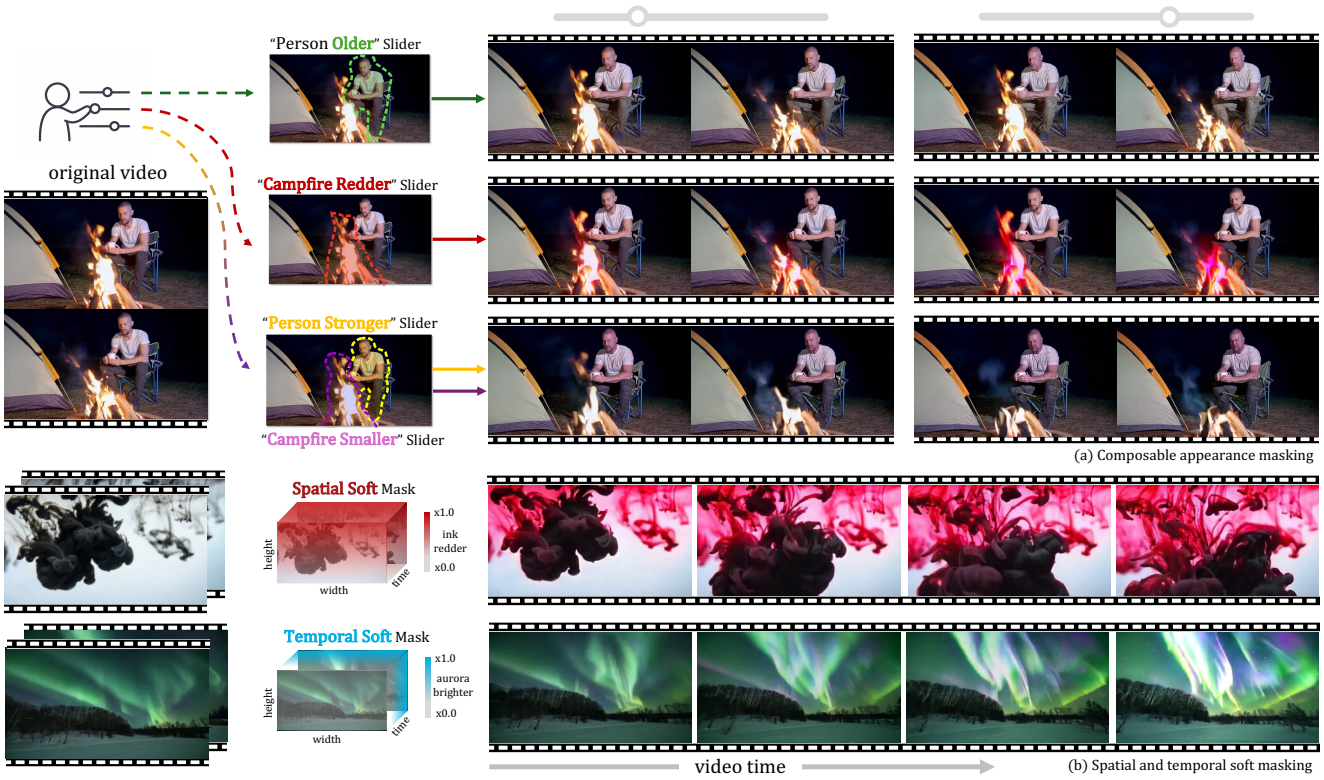

Figure 2. **Explicit spatiotemporal masking.** (top) *Composable masking:* localize different sliders to different concepts (person vs. campfire) and compose them in one video. (bottom) *Spatial/temporal masking:* leveraging *TokenDial* and soft masks, we can easily make only the top portion of the ink videos redder and create a gradient effect, or make aurora brighter only towards the end of the video.

real creative workflows, realism is only the starting point: creators need the ability to keep a scene fixed while precisely controlling how much a particular attribute changes. Prompts can specify what appears (e.g., "a campfire", "a person"), but they are a weak interface for dialing attribute strength (e.g., fire intensity, aging, or motion magnitude) without drifting identity, background, or temporal coherence. We study continuous semantic scalar control: slider-style modulation of continuous attributes while preserving the rest of the video.

A major missing capability is motion dynamics control. Existing "video sliders" primarily target appearance attributes, while continuously dialing motion properties, such as intensity, rhythm, or perceived speed, remains brittle. Equally importantly, current controls are largely implicit in space and time: whether a learned edit takes effect at a particular region or moment is often decided by the model, rather than specified by the user. Table 1 highlights this gap: prior methods do not provide reliable motion-dynamics sliders, nor an explicit spatiotemporal interface to determine where and when an edit applies.

Prior approaches only partially address these needs. One way is to use progressive prompts to edit the video via video-to-video (V2V) models. While they can change video content, these models do not offer continuous, monotonic control of attribute strength required for slider-style manipulation. Slider methods based on discrete categories or lightweight weight updates (e.g., LoRA [14]) can learn an appearance editing direction, yet they can be training-intensive and difficult to disentangle and compose. In addition, low-rank adaptors and other finetuning-based methods alter the base model weights, which risks degrading the general capability of the model and can result in overfitting. More importantly, because these finetuned models now have updates baked into the weights, making localized

Table 1. Capabilities of baselines.

Method	Appear.&. Motion Attr.	Prog. Scaling	Explicit. Mask.	ID. preserve.
I2I + I2V	✗	✗	✓	✓
Text-based V2V	✗	✗	✓	✓
Concept Sliders	✗	✓	✗	✗
SliderSpace	✗	✓	✗	✗
Text Slider	✗	✓	✗	✗
FreeSliders	✗	✓	✗	✗
***TokenDial* (Ours)**	✓	✓	✓	✓

edits remains difficult (e.g., only make the left person older but not the right).

We present *TokenDial* built on the following observation: the visual patch token space in pretrained T2V diffusion transformer (DiT; [30]) models contains directions that correspond to semantic attribute changes. We show that by leveraging the strong priors of a pretrained understanding model, a single offset vector in the token space can be learned to associate to a specific attribute; when these offsets are added to the visual tokens, they can strengthen or weaken the associated attributes. This not only provides a straightforward mechanism to scale the effects of a desired attribute or compose multiple attributes, it also provides an intuitive interface to control how attributes are changed spatiotemporally by adding to specific patch tokens during denoising. This controllability makes *TokenDial* naturally suited for fine-grained controls, including motion dynamics that were not previously addressed by existing slider-based methods.

We validate *TokenDial* through extensive experiments on diverse attributes for both appearance (e.g., aging, weather) and motion (e.g., higher/lower motion magnitude). Quantitative benchmarks and human evaluations confirm that *TokenDial* achieves superior controllability and editing quality compared to state-of-the-art baselines, effectively balancing attribute modification with content preservation.

2. Related Work

Controllable video generation. Recent advances have enabled diverse forms of control in generative models, ranging from architecture- and conditioning-based designs [2, 7, 44] to application-driven controls such as subject personalization [11, 15, 33] and camera motion guidance [1, 12, 22, 26, 41]. While these directions greatly improve usability, they typically address discrete or coarse controls (what to generate, where to place it, or which trajectory to follow), rather than continuously dialing the strength of an attribute during generation. In contrast, we study continuous, slider-style control of fine-grained appearance and motion attributes in T2V models.

Video editing. Instruction- or example-based video editing methods [17, 20, 23, 36, 40, 48] achieve strong open-domain edits. However, their control interface is typically discrete: edit strength is specified through instruction text or a small set of pre-defined levels, which makes it difficult to achieve continuous, monotonic slider-style modulation of attribute strength. They also often rely on substantial edited supervision or synthetic training data, limiting scalability. In contrast, we target continuous attribute control during generation, using pretrained understanding signals rather than large-scale paired edits.

Fine-grained and continuous video generation. This line of work explores slider-based controls on attributes with various setups and architectures [4, 6, 9, 10]. FreeSliders [6] and Text Slider [4] have begun to study slider-style control for videos. FreeSliders provides training-free concept steering via an inference-time approximation of noise updates, whereas Text Slider learns lightweight LoRA directions in the shared text encoder for plug-and-play, reusable control. Other work in the image domain also pursue fine-grained, continuous edits by injecting LoRA modules or steering text/conditioning tokens [9, 10, 18, 28]. Despite controllability, existing sliders are largely appearance-centric and keep the spatiotemporal extent of edits implicit, which limits its precise localization, weakens identity preservation, and makes motion dynamics difficult to dial.

Semantics of various latent spaces. Semantic structure in representation spaces has long enabled controllable generation, from linear directions in GAN latent space [3, 16, 31] to semantic bottlenecks in diffusion models that support editing and interpretability [5, 21, 29]. Recent work further shows that aligning denoising representations with pretrained understanding models improves training efficiency and generation quality, including extensions to video [34, 43, 46]. TokenVerse [11] revealed the modulation space in Flux-like models can be used for personalization of diverse concepts. Inspired by these findings, we propose to explore the token embedding space of video DiTs to augment specific attributes using external understanding models.

3. Method

In this section, we detail our method formulation that leverages $\mathcal{V}^+$ space for semantic (appearance and motion) attribute scaling. We first formalize the learnable space and define the offset vector with which we align the attribute semantics to (§3.1); we then introduce the self-supervised training objectives (§3.2). Finally we describe the inference set up using our method (§3.3). The overview of our method is illustrated in Figure 3.

During inference, a pretrained T2V diffusion model generates a sample by integrating a learned vector field in continuous time. We denote x_t as the latent video at time $t \in [0, 1]$, c as the conditioning input (e.g., text prompts), and θ as the frozen model parameters. Starting from Gaussian noise $x_1 \sim \mathcal{N}(0, I)$, sampling follows the ODE: $\frac{dx_t}{dt} = \Phi_\theta(x_t, t, c)$, which is integrated from $t = 1$ to $t = 0$ to obtain the generated video x_0. During training, we sample a random time t and train the model to predict the denoising direction (i.e., the vector field) from x_t under conditioning c, via the flow-matching objective.

3.1. Construct the learnable token offsets

Visual patch-token space $\mathcal{V}$. A pretrained video DiT operates on a sequence of visual patch tokens. Given a la-

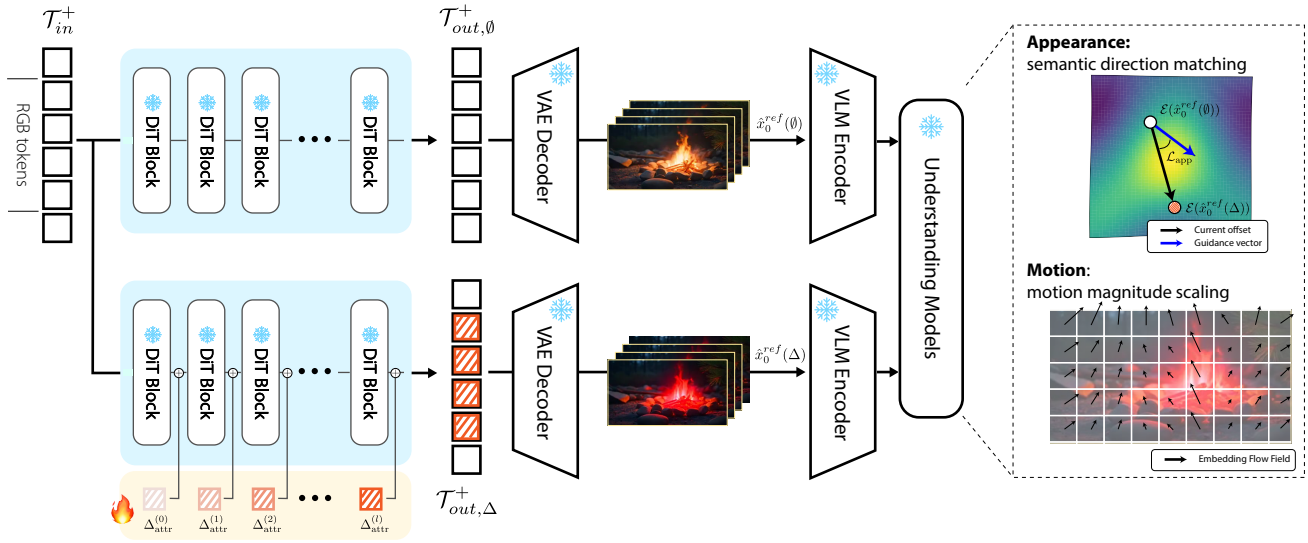

Figure 3. **Overview of *TokenDial*.** We inject learnable spatiotemporal token offsets into intermediate video patch tokens of a frozen text-to-video DiT. Offsets are trained with external understanding models: appearance via semantic direction matching and motion via motion-magnitude scaling.

tent video $x \in \mathbb{R}^{C\times F\times H\times W}$ (encoded by a VAE/video encoder), the model applies a patchification operator $\mathrm{Patchify}(\cdot)$ to obtain

$$\mathcal{T} = \mathrm{Patchify}(x) \in \mathbb{R}^{L\times d}, \tag{1}$$

where d is the hidden dimension and L is the number of visual patches (tokens), which depends on the video resolution and length. We denote the set of such patch-token sequences by $\mathcal{V} := \mathbb{R}^{L\times d}$, and a token sequence as $\mathcal{T} \in \mathcal{V}$. Importantly, patchification preserves the spatial-temporal correspondence: each token index $i \in \{1,\dots,L\}$ maps to a specific region in space and time, providing an explicit handle to address *where* and *when* an edit is applied.

Offset space $\mathcal{V}^+$. We introduce an offset space $\mathcal{V}^+ := \mathbb{R}^d$, whose elements are additive offset vectors $\Delta \in \mathbb{R}^d$ that are applied to visual patch tokens. Since $\mathcal{V}^+$ depends only on d, it is agnostic to video resolution and length. Applying an offset uniformly (or selectively) to tokens yields a new token sequence

$$\mathcal{T}' = \{\, t_i + s_i\Delta \mid t_i \in \mathbb{R}^d,\ i = 1,\dots,L \,\}, \tag{2}$$

where $s = \{s_i \in [0,1]\}$ is a (soft) spatiotemporal mask. This formulation makes control explicit: by choosing s, we can specify which regions/frames are affected, and by scaling Δ we obtain a linear strength dial once a direction is learned.

We further allow the offset to be layer-dependent. Let $\Delta^{(k)} \in \mathbb{R}^d$ denote the offset applied after the k-th DiT block, leading to

$$\mathcal{T}'^{(k)} = \{\, t_i + s_i\Delta^{(k)} \mid t_i \in \mathbb{R}^d,\ i = 1,\dots,L \,\}. \tag{3}$$

This adds a small number of learnable degrees of freedom without modifying the backbone weights (see Figure 3 for an overview). For example, in our implementation, the learnable offsets $\{\Delta^{(k)}\}$ introduce only 0.256% as many trainable parameters as a rank-64 LoRA. Note that unlike textual inversion [8], this does not alter the token length nor the attention structure.

3.2. Embed semantics onto token offsets

Our goal is to align the token offset with a target attribute semantics. We therefore learn an attribute-specific offset vector

$$\Delta_{\text{attr}} \in \mathcal{V}^+ , \tag{4}$$

and apply it to the visual patch tokens as described in Sec. 3.1. When layer indices are not essential, we drop the superscript (k) for clarity. We optimize Δ_{attr} while keeping the pretrained generator frozen, using gradient feedback from external pretrained understanding models (Fig. 3). We next describe a stable training procedure.

Multi-step posterior refinement. As shown in Figure 3, all components except for the offset vectors are frozen during training. A naive objective would apply the external supervision directly on the one-step estimate videos obtained at an intermediate noise level. At each training iteration we sample a timestep t and obtain x_t from the clean latent x_0 via the forward noising process. With the backbone frozen, we run the denoiser twice: (i) without offsets, and (ii) with offsets injected into the token stream, producing two posterior estimates of the clean latent, denoted by $\hat{x}_0(\varnothing)$ and $\hat{x}_0(\Delta)$ (via Tweedie's formula [13]). However, $\hat{x}_0$ recon-

structed from a highly noisy x_t lacks high-frequency details, which makes gradients from understanding models unstable and noisy.

To stabilize training, we refine the posterior estimate by unrolling a small number of additional denoising steps before computing the loss. Specifically, starting from the initial prediction at timestep t, we run K extra reverse steps to obtain a refined estimate $\hat{x}_0^{\text{ref}}$. For efficiency, we stop gradients through the refinement unroll and backpropagate only through the initial prediction, similar to [24]. In our experiments we use a small K (e.g., K=4) for high-noise timesteps.

Appearance Control: Semantic Direction Matching. We supervise appearance control in a pretrained video-understanding embedding space using InternVideo2 [39]. Let $\mathcal{E}(\cdot)$ denote the InternVideo2 visual encoder. Given refined reconstructions with and without offsets, we define the predicted attribute direction in the embedding space as

$$\mathbf{d}_{\text{pred}} = \mathcal{E}\big(\hat{x}_0^{\text{ref}}(\Delta)\big) - \mathcal{E}\big(\hat{x}_0^{\text{ref}}(\varnothing)\big)\,. \tag{5}$$

We align this direction with a target direction $\mathbf{d}_{\text{tgt}}$ using cosine distance:

$$\mathcal{L}_{\text{app}} = 1 - \cos(\mathbf{d}_{\text{pred}}, \mathbf{d}_{\text{tgt}})\,. \tag{6}$$

To preserve identity and scene content, we add a perceptual regularizer based on LPIPS [45] between two reconstructions. The final appearance objective is

$$\mathcal{L}_{\text{appear}} = \mathcal{L}_{\text{app}} + \lambda_a \cdot \text{LPIPS}\big(\hat{x}_0^{\text{ref}}(\Delta), \hat{x}_0^{\text{ref}}(\varnothing)\big)\,, \tag{7}$$

where λ_a controls the regularization strength. This objective encourages the offset to realize the desired attribute change while minimally disturbing the underlying content. We obtain $\mathbf{d}_{\text{tgt}}$ by contrasting either paired text prompts (e.g., "hot" vs. "cold") or exemplar videos, and projecting them into the same InternVideo2 feature space.

Motion Control: Motion Magnitude Scaling. We next supervise motion dynamics (e.g., making an action higher/lower motion magnitude or more/less intense). While video foundation model embeddings can capture temporal information, we found them less reliable for quantifying motion strength during training, as the resulting signals are sensitive to the keyframe sampling.

Instead, we directly measure motion magnitude in a feature space that is stable and semantically meaningful. We extract frame-wise patch embeddings using DINOv2 [27] and compute optical flow on these embeddings with the Lucas–Kanade (LK) method [25], yielding a motion field $\mathbf{m}(\cdot)$. Our goal is to scale motion strength by a factor γ ($\gamma > 1$ amplifies motion; $\gamma < 1$ attenuates it). Crucially, we use a self-supervised target: using a fixed reference video (e.g., the no-offset prediction) becomes misaligned after training, since the offset video can evolve faster/slower and thus no longer matches the reference frame-to-frame. In this case, optical flow comparisons across two sequences produce inconsistent correspondences and high-variance gradients. We therefore scale the motion field of the offset sample itself and use stop-gradient $.\text{sg}()$ to define a stable target:

$$\mathcal{L}_{mot} = \|\mathbf{m}(\hat{x}_0^{ref}(\Delta)) - \gamma \cdot [\mathbf{m}(\hat{x}_0^{ref}(\Delta)).\text{sg}(\,)]\|_2^2. \tag{8}$$

DINOv2 patch features provide temporally stable local anchors, reducing spurious jitter when scaling motion magnitude. To better preserve identity and scene content while changing dynamics, we encourage the first frame to remain consistent across attribute scales by regularizing DINOv2 features on the first frame:

$$\mathcal{L}_{\text{ff}} = 1 - \cos\Big(\mathcal{D}\big(\hat{x}^{\text{ref}}_{0,\,t=0}(\Delta)\big)\,, \mathcal{D}(\hat{x}^{ref}_{0,t=0}(\Delta).\text{sg}(\,)\Big)\,, \tag{9}$$

where $\mathcal{D}(\cdot)$ denotes the DINOv2 feature extractor on a single frame.

Finally, the motion objective is

$$\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{mot}} + \lambda_m\, \mathcal{L}_{\text{ff}}, \tag{10}$$

where λ_m controls the regularization strength. Minimizing this objective learns offsets that modulate the magnitude of motion while preserving the underlying motion pattern and identity cues.

3.3. Inference with Token Offset

After training, the learned offset Δ acts as an attribute direction in token space: injecting it into the generator produces the desired semantic change. As shown in Figure 9e, simply injecting this offset globally already yields the desired attribute change (e.g., making the campfire bluer while preserving the tent). However, to achieve precise spatiotemporal localization, strictly preserve the background (Figure 9f), and achieve the slider effect (Figure 1), we expose two complementary controls: *where*/*when* the edit applies and how strongly it is applied.

Structure-Aware Spatiotemporal Modulation. We observe that Δ specifies *what* to change, while the model's attention maps reveal *where* and *when* the change should occur. Early in denoising, we extract attention from the target text token (e.g., "campfire") to visual patch tokens and aggregate it into a token-level soft mask $\mathbf{s} \in [0,1]^L$, where L is the number of spatiotemporal patch tokens. Although $\mathbf{s}$ is defined over tokens, patchification preserves spatiotemporal correspondence, so $\mathbf{s}$ can be viewed as a soft matte over space and time. We gate the offsets using $\mathbf{s}$, confining the edit to the object's trajectory across time and protecting unrelated regions from drift. Importantly, the mask is soft rather than binary, enabling subsequent self-attention to

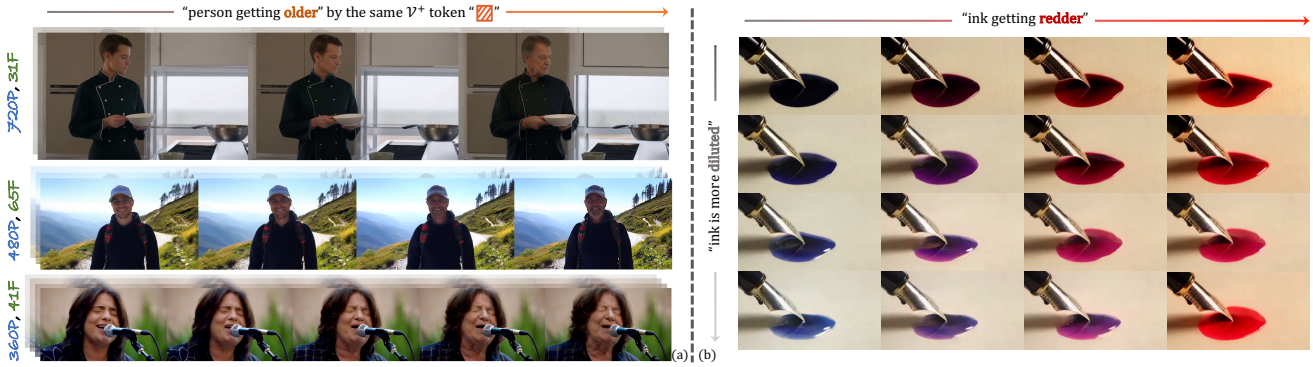

Figure 4. **(a)** *TokenDial* learned token offsets transfers zero-shot across video resolutions and lengths. **(b)** *TokenDial* composes attributes by combining offsets, enabling independent control along multiple sliders (e.g., ink "redder" and "more diluted").

propagate the change to correlated regions when appropriate. For instance, in Fig. 2b, as the aurora becomes brighter, the mountain region also brightens due to the reflected illumination, while the underlying identity and structure are preserved.

Intensity Control via Compositional Flow Guidance. To control the magnitude of the edit (the "slider" effect), a naive choice is to scale the offset Δ directly in feature space, which can distort the generation trajectory at large magnitudes. Instead, we dial edit strength at the vector field level by composing a text-consistent base flow with an edit-induced flow. Let $\Phi_\theta(x_t, t, c)$ denote the predicted vector field and $\varnothing$ the unconditional input. We define the update $\tilde{u}$ as

$$\begin{aligned}\tilde{u} = \Phi_\theta(x_t, \varnothing) + s_{\text{txt}}(\Phi_\theta(x_t, c) - \Phi_\theta(x_t, \varnothing)) \\ + s_{\text{edit}}(\Phi_\theta(x_t, c, \Delta) - \Phi_\theta(x_t, c))\,.\end{aligned} \tag{11}$$

The first two terms form the standard text guidance and define the base trajectory, while the third term isolates the differential velocity induced by Δ. Adjusting s_{edit} scales only the edit-induced component, enabling continuous slider control while preserving the text-consistent structure.

4. Applications

TokenDial provides a practical control interface for T2V generation that goes beyond existing video sliders: it supports continuous modulation of motion dynamics, explicit where/when localization with composable edits, and strong transfer across video settings and model architectures. We highlight these below.

Continuous control of appearance and motion dynamics. *TokenDial* enables slider-style control over a broad set of attributes, spanning both appearance (e.g., aging, color, lighting) and motion dynamics. Varying the offset strength yields smooth, monotonic changes while largely preserving identity, background, and temporal coherence, shown in Figure 1 and Figure 7.

Explicit spatiotemporal locality and compositional edits. Offsets specify *what* to change, while attention-derived masks or user-provided masks specify *where* and *when* the change applies. This produces a soft spatiotemporal matte that confines edits to the target trajectory and protects unrelated regions from drift (Figure 2). Because offsets can be gated independently, *TokenDial* naturally supports composition: multiple regions (Figure 2a) and multiple attributes (Figure 4b) can be edited in the same video without interfering with each other.

Generalization across video length and resolution. Offsets live in $\mathcal{V}^+$ and depend only on the hidden dimension, making them agnostic to the number of spatiotemporal tokens. As a result, an offset trained on shorter, lower-resolution clips transfers zero-shot to different video lengths and resolutions supported by the backbone, enabling efficient training while retaining high-resolution controllability at test time. Figure 4a shows unified "person older" token offsets generalized to different video length and resolution at test time.

Generalization across video model architectures. *TokenDial* is not tied to a specific backbone design. Our main experiments use an internal DiT backbone that relies on full self-attention and does not provide an explicit modulation space, similar to [42, 47]. We demonstrate that the same token-offset principle transfers to Wan 2.1 [37], a cross-/self-attention-based video model equipped with modulation layers, Fig. 5 and Fig. 4a show the results generated by Wan. To apply *TokenDial* on Wan, we define the injection point in its feature stream at the self-attention residual and learn offsets in that space, obtaining consistent controllability. This cross-architecture transfer indicates that *TokenDial* captures a general mechanism for continuous control instead of being model-specific.

Semantic disentanglement. Fig. 6(b) shows that supervision in the InternVideo2 [39] embedding space can inherit

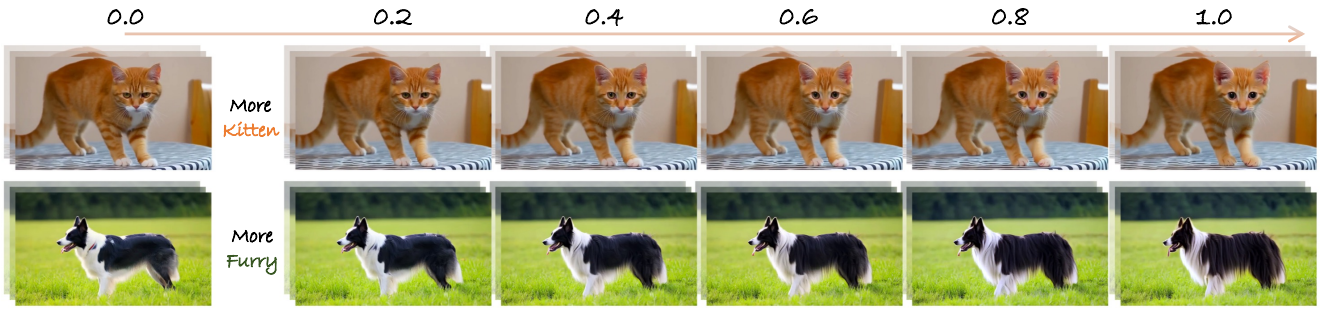

Figure 5. **Generalization to Wan**. *TokenDial* transfers to the Wan 2.1 backbone, enabling continuous appearance sliders by injecting offsets into Wan's feature stream. Examples show a "more kitten" slider (cat) and a "more furry" slider (dog).

spurious correlations (e.g., the direction for "older" may unintentionally co-vary with body weight). We mitigate this by projecting out biased *principal* directions in the supervision space, following a simple debiasing strategy similar to [35], which yields better disentangled edits (Fig. 6).

(a) original (b) biased edit (c) de-biased edit

Figure 6. **Semantic debiasing.** "Older" edits learned from InternVideo2 can also increase body weight (b); debiasing removes this coupling (c).

5. Experimental Setup

Concepts, attributes, and slider tasks. We study slider-style control defined over a concept-attribute pair. A *concept* is the main entity or phenomenon described in the prompt (e.g., campfire, ink, person) with localized spatiotemporal support in the video. An *attribute* is a continuous scalar property of that concept that can be modulated while preserving scene layout and identity (e.g., color intensity, size, dilution, motion magnitude). A *slider task* evaluates whether a method can produce progressive, monotonic changes along the attribute scale under the same prompt and seed. We consider two families of tasks: appearance sliders modulate object-centric spatial attributes under strict localization, with background and identity preserved; motion sliders modulate motion magnitude while preserving the underlying motion pattern.

***TokenDial* dataset and evaluation protocol.** We evaluate on 12 concepts spanning particle systems, volumetric phenomena, fluids, and articulated subjects, with 5 attributes per concept. Appearance offsets are trained on small, concept-specific video–text collections without paired edits. Motion offsets are trained on a few hundred green-screen clips and applied universally across prompts during inference. See the supplementary material for dataset details. For each concept-attribute pair, we generate 16 base videos and evaluate 5 slider strengths under identical prompts, seeds, and inference budgets, resulting in roughly 4,500 videos per method. We additionally show that *TokenDial* can be trained using synthetic text-video pairs generated by the backbone itself, while retaining similar controllability; i.e., the same "person older" slider can be learned from real videos (Fig. 2a) or from synthetic videos generated by the model (Fig. 4a).

Implementation details. We optimize token offsets on a frozen pretrained text-to-video DiT model, sharing a similar architecture to [42, 47]. Unless specified otherwise, we use λ_a=0.5 for appearance and λ_m=5.0 for motion, and train offsets for 300 steps with AdamW (lr $= 10^{-5}$). Training uses 32-frame videos at 24 FPS with 320×176 resolution; at inference we broadcast offsets to longer sequences (up to 64 frames). We use classifier-free guidance with text scale s_{txt}=4.5 and vary the edit strength with $s_{\text{edit}} \in [0, 1]$ across all experiments.

Baselines. We compare against representative methods for continuous control. Our primary video slider baselines are FreeSliders [6] and Text Slider [4]. We also adapt image-domain sliders, Concept Sliders [9] and SliderSpace [10], to videos by applying their learned directions consistently across frames. Finally, we include a strong text-driven video editing baseline, UniVideo [40], by progressively strengthening the instruction (e.g., *slightly* $\rightarrow$ *moderately* $\rightarrow$ *much* $\rightarrow$ *extremely*), and an I2I+I2V pipeline that combines the image slider Kontinuous Kontext [28] with Senorita [48] for propagating first-frame edits.

6. Results

6.1. Qualitative Comparison

In Fig. 7, we qualitatively compare *TokenDial* with prior slider-based methods. For appearance control, Text Slider often shows a weak response, while FreeSliders and image-

Figure 7. **Qualitative comparison on appearance and motion sliders**. *TokenDial* achieves smooth, continuous slider control for appearance (a) and motion dynamics (b), with stronger edits and better preservation than prior methods.

domain sliders adapted to video (Concept Sliders, Slider-Space) frequently introduce identity drift or background changes as edit strength increases. The I2I+I2V pipeline is limited by first-frame editing: the image slider affects only the initial frame, and the propagated edit can miss objects that appear later in the video. Text-driven V2V editing does not provide reliable progressive strength control via prompts alone. In contrast, *TokenDial* produces strong and continuous appearance edits while better preserving identity and context.

For motion dynamics, competing methods show limited ability to amplify motion magnitude. *TokenDial* successfully scales dynamics (e.g., walking $\rightarrow$ running), producing larger pose, clothing, and hair displacements while maintaining coherent structure.

6.2. Quantitative Comparison

We evaluate *TokenDial* from two complementary perspectives: (i) slider controllability: whether edit strength changes progressively, smoothly, and monotonically; and

Table 2. Quantitative comparison on VLM metrics and video quality metrics.

Method	VLM Evaluation					Video Qual.	Text Align.
	Edit Qual. ↑	ID Pre. ↑	BG Pre. ↑	Continuity ↑	Overall ↑	PickScore ↑	ViCLIP ↑
ConceptSlider	3.635	4.722	4.599	3.654	4.153	19.610	24.737
SliderSpace	3.743	4.722	4.603	3.738	4.202	19.659	24.813
FreeSlider	3.932	4.830	4.775	3.936	4.368	19.697	**25.242**
TextSlider	2.771	4.970	**4.974**	2.833	3.887	**19.726**	24.936
Kontinuous Kontext +Senorita (I2I+I2V)	2.578	4.868	4.940	2.518	3.726	19.567	24.353
UniVideo (V2V)	3.032	4.754	4.874	2.786	3.862	19.461	23.722
Ours	**4.165**	**4.988**	4.959	**4.234**	**4.587**	19.651	24.942

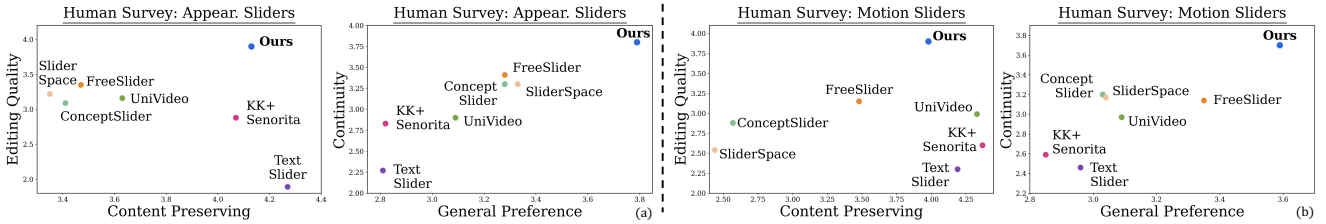

Figure 8. **Human survey results.** We compare methods on appearance (a) and motion sliders (b) using quadrant plots of edit quality vs. content preservation, and slider continuity vs. video general preference.

Table 3. Quantitative evaluation of slider controllability.

Method	Editing Quality			Preserve	
	Range(CR) ↑	Mono. ↑	Smooth(CSM) ↓	LPIPS(SP) ↓	Overall ↑
ConceptSlider	0.210	0.512	0.735	0.391	0.416
SliderSpace	0.451	0.550	0.676	0.403	0.645
FreeSlider	0.378	0.567	0.537	0.296	0.755
TextSlider	0.004	0.501	**0.262**	0.110	0.742
Slider-based I2I+I2V	0.196	0.542	0.373	**0.085**	0.808
Text-based V2V	0.276	0.534	1.154	0.440	0.037
Ours	**0.460**	**0.573**	0.386	0.249	**0.982**

(ii) edit quality and preservation: whether the edit is correct while identity, background, and temporal coherence are maintained.

Slider controllability. Following FreeSliders [6], we report Conceptual Range (CR), Conceptual Smoothness (CSM), Semantic Preservation (SP), and a monotonicity score. CR measures the semantic span of a slider using CLIP [32] distance between endpoint generations. CSM measures how uniformly CLIP scores change across strength levels (lower is better), and monotonicity measures whether the semantic change progresses consistently in one direction. SP measures content preservation across slider levels using LPIPS [45]. A key caveat is that CSM/SP can favor conservative methods: approaches that make very small edits may appear smooth and preserving, yet have limited semantic range (low CR). This behavior is visible for Text Slider and the I2I+I2V pipeline, which achieve strong preservation scores but weak semantic response in Fig. 7. To balance strength and stability, we follow [6] to report the overall score (OS):

$$\text{OS} = \frac{\text{CR}}{\epsilon + \text{SP}} + (1 - \text{CSM}), \tag{12}$$

where ϵ is set to 1 following [6]. As shown in Table 3, *TokenDial* achieves the best OS (0.982), far above conservative baselines with high CSM/SP but weak edits (I2I+I2V: 0.808; Text Slider: 0.742), indicating that it attains large semantic range while maintaining smooth, stable transitions and strong preservation. Notably, prompt-driven V2V editing (UniVideo) exhibits less consistent progression across strength prompts, as shown in Fig. 7, and correspondingly attains the highest (worst) CSM score, indicating poor editing continuity.

VLM-based evaluation. Following EditVerse [17], we further assess editing quality, identity preservation, background preservation, and temporal continuity using a VLM-based rubric. As shown in Table 2, *TokenDial* obtains the highest scores on editing quality, identity preservation, and continuity, while maintaining comparable video quality and text alignment measured by PickScore [19] and ViCLIP [38]. Notably, methods with high preservation scores but weak edits (e.g., Text Slider and the I2I+I2V pipeline) align with the qualitative comparison in Fig. 7, where attribute changes are visibly limited.

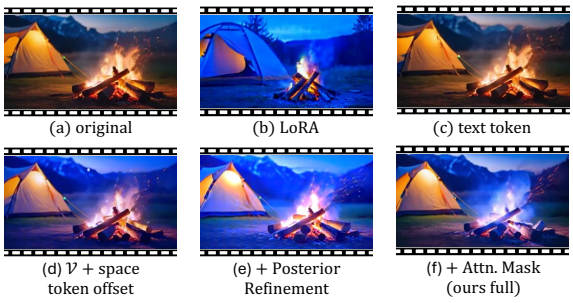

Figure 9. **Ablation**. (b) and (c) show the alternative to token offset, and (d) to (e) shows our results as additional components are progressively integrated.

Table 4. Ablation analysis of components for controllable editing.

	Editing Quality			Preserve	
Method	**Range ↑**	**Mono. ↑**	**Smooth ↓**	**LPIPS ↓**	**Overall ↑**
Text Token	0.011	0.458	**0.241**	**0.129**	0.769
LoRA	0.393	0.550	0.758	0.434	0.516
$\mathcal{V}^+$-Space Offset	0.412	0.566	0.419	0.270	0.905
+ Posterior Refine	**0.499**	0.570	0.440	0.298	0.944
+ Attn. Mask (ours full)	0.460	**0.573**	0.386	0.249	**0.982**

6.3. Human Evaluation Study

We conduct a human study to assess edit effectiveness and preservation, with an emphasis on motion dynamics, which are difficult to evaluate reliably with existing metrics and keyframe-based VLM scores. We randomly sample 32 appearance sliders and 32 motion sliders, and recruit 212 participants from Prolific. In each trial, participants view the original video followed by edited outputs at increasing slider strengths. They rate each method on four 0–5 Likert scales: (1) prompt-following/edit quality, (2) identity/content preservation, (3) temporal continuity, and (4) overall preference. Each participant evaluates *TokenDial* and all baselines in a randomized order for fair comparison.

As summarized in the quadrant plots (Fig. 8), *TokenDial* is clearly preferred for motion edits (Fig. 8b), achieving the highest perceived edit quality while maintaining strong preservation. Competing methods either fail to produce noticeable temporal changes or degrade coherence when attempting stronger motion edits. For appearance edits (Fig. 8a), human preferences closely match our quantitative results, confirming that *TokenDial* delivers strong edits while preserving identity and context. See the supplementary material for the full protocol and additional human study analyses.

6.4. Ablation study

Fig. 9 isolates the key design choices of *TokenDial*. Starting from a minimal baseline, we progressively add each component to reveal its role in controllable and localized editing.

Offset parameterization. We first replace our $\mathcal{V}^+$ token offsets with alternative parameterizations, including LoRA-based tuning and text-token tuning. As shown in Fig. 9b, LoRA-based edits tend to affect the entire video (including background) due to the lack of an explicit spatial handle, whereas text-token tuning yields negligible visual change (Fig. 9c).

$\mathcal{V}^+$-space offsets. Introducing additive offsets in $\mathcal{V}^+$ provides a direct control handle on visual patch tokens, enabling localized appearance edits while better preserving background content (Fig. 9d).

Posterior refinement. Without refinement, supervision from understanding models can be unstable at high noise levels, leading to weak edits. Multi-step posterior refinement stabilizes training and yields semantically meaningful changes (Fig. 9e).

Attention-derived masking. Finally, attention-based masking further confines edits to the target concept across space and time, improving focus and reducing background drift (Fig. 9f). Table 4 corroborates these observations, showing consistent improvements as each component is added.

7. Conclusion

In this work, we propose *TokenDial* that introduces continuous slider control to pretrained text-to-video models by learning additive token offsets in an intermediate spatiotemporal patch-token space. We show that this offset space is semantic: scaling a learned direction yields smooth, predictable changes in attribute strength, enabling controllable edits for both appearance and motion dynamics without retraining the backbone. More broadly, our efforts open avenues for interactive, fine-grained controllable video generation and richer semantic control in future generative systems.

Limitations. Our method relies on pretrained video understanding models (e.g., InternVideo2) to define semantic directions for appearance control. While effective in many cases, the underlying understanding space may exhibit entanglement and biases that are difficult to fully disentangle. Beyond high-level semantic biases (e.g., age correlated with body weight), we observe that certain low-level attributes, such as color changes, can be entangled with other visual factors in the embedding space. In such cases, simple semantic debiasing via subspace projection may be insufficient to isolate the desired attribute without affecting related properties.

References

[1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video, 2025. 3

[2] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-{\delta}: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 3

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016. 3

[4] Pin-Yen Chiu, I-Sheng Fang, and Jun-Cheng Chen. Text slider: Efficient and plug-and-play continuous concept control for image/video synthesis via lora adapters, 2025. 3, 7, 13

[5] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models, 2023. 3

[6] Rotem Ezra, Hedi Zisling, Nimrod Berman, Ilan Naiman, Alexey Gorkor, Liran Nochumsohn, Eliya Nachmani, and Omri Azencot. Freesliders: Training-free, modality-agnostic concept sliders for fine-grained diffusion control in images, audio, and video. *arXiv preprint arXiv:2511.00103*, 2025. 3, 7, 9, 13

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 3

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 4

[9] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2024. 3, 7, 13

[10] Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15994–16003, 2025. 3, 7, 13

[11] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space, 2025. 3

[12] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 4

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[15] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation, 2025. 3

[16] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020. 3

[17] Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, Daniil Pakhomov, Zhe Lin, Soo Ye Kim, and Qiang Xu. Editverse: Unifying image and video editing and generation with in-context learning, 2025. 3, 9

[18] Ronen Kamenetsky, Sara Dorfman, Daniel Garibi, Roni Paiss, Or Patashnik, and Daniel Cohen-Or. Saedit: Token-level control for continuous image editing via sparse autoencoder. *arXiv preprint arXiv:2510.05081*, 2025. 3

[19] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 9

[20] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks, 2024. 3

[21] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023. 3

[22] Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li. Generative video motion editing with 3d point tracks. *arXiv preprint arXiv:2512.02015*, 2025. 3

[23] Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, and Jiaya Jia. Generative video propagation. *arXiv preprint arXiv:2412.19761*, 2024. 3

[24] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10822–10832, 2024. 5

[25] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981. 5

[26] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control, 2024. 3

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 5, 17

[28] Rishubh Parihar, Or Patashnik, Daniil Ostashev, R. Venkatesh Babu, Daniel Cohen-Or, and Kuan-Chieh Wang. Kontinuous kontext: Continuous strength control for instruction-based image editing, 2025. 3, 7, 13, 15

[29] Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models, 2023. 3

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[31] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 3

[32] 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. *CoRR*, abs/2103.00020, 2021. 9

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 3

[34] Wonyong Seo, Jaeho Moon, Jaehyup Lee, Soo Ye Kim, and Munchurl Kim. Propfly: Learning to propagate via on-the-fly supervision from pre-trained video diffusion models, 2026. 3

[35] Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, and Garrison W. Cottrell. Discovering and mitigating biases in clip-based image editing. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2972–2981, 2024. 7

[36] DecartAI Team. Lucy edit: Open-weight text-guided video editing. 2025. 3

[37] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. 6, 19

[38] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 9

[39] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024. 5, 6, 17

[40] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025. 3, 7, 13, 16

[41] Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation, 2025. 3

[42] 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. 6, 7

[43] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think, 2025. 3

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[45] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 5, 9

[46] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models, 2025. 3

[47] 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, 2024. 6, 7

[48] Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. In *NeurIPS D&B*, 2025. 3, 7, 13, 16

Appendix

In this supplementary material, we provide additional details and results to support the main paper. We include human evaluation details (§A), descriptions of the slider-based I2I-I2V and text-based V2V baselines (§B), dataset details (§C), an analysis of posterior refinement (§D), Lucas–Kanade optical flow on DINOv2 patch features (§E), implementation details on Wan 2.1 (§F), the VLM prompt used for slider evaluation (§G), and additional qualitative results (§H).

Contents

A. Human Evaluation Study

We conducted human evaluation study using a Qualtrics survey distributed through Prolific. As shown in Fig. 11, each survey question presented an original video together with four edited videos. The edited videos were generated by one of six baseline methods or by our method, chosen randomly, and with edit strength increasing progressively across the four edits. The baselines included Concept-Slider [9], SliderSpace [10], FreeSlider [6], Text Slider [4], Kontinuous Kontext [28] combined with Senorita [48] as a slider-based image-to-image (I2I) plus first-frame propagation video editing baseline, and UniVideo [40] as a text-based video-to-video (V2V) editing baseline.

Below the videos, participants rated six statements on a Likert scale. These questions were aligned with our VLM-based evaluation and were designed to measure (1) edit quality and progression, (2) content preservation, and (3) subjective human preference. On each survey page, participants were shown five sets of video edits, one from each method being compared, together with the corresponding questions. This design ensured that every participant evaluated all compared baselines the same number of times, enabling a fair comparison across methods. The order of the baselines was randomized.

Table 5. Mapping our evaluation metrics to the Likert statements in our study

Metric	Question
Editing Quality	The **attribute** becomes **modifier** in the edited videos.
ID Preserve	Despite the changes, the people/objects are clearly recognizable as the **same people/objects**.
BG Preserve	The parts of the video that are **not related to the attribute** remained unchanged in the edited videos.
Continuity	The intensity of the edit increases **progressively** from the first video to the last (left to right).
Artifacts	The edited videos **do not add artifacts** (e.g., glitches, grid-like patterns, unrealistic visuals).
Useful	These **edits provide a reasonable range of the effects** that I can choose from in different editing scenarios.

We generated a total of 64 editing directions for this study, corresponding to real-world video editing tasks such as making a person walk faster or making a fire brighter. Among them, 32 editing directions targeted appearance attributes, while the other 32 targeted temporal or dynamic changes. As shown in Fig. 11, the editing prompt was inserted directly into the survey; for example, *campfire/smaller* could be replaced with *speed/slower*.

In total, we recruited **212 participants** through Prolific. We first compared *TokenDial* against video-slider-based methods. Specifically, FreeSlider [6] and Text Slider [4] are originally designed for video slider control, while Concept-Slider [9] and SliderSpace [10] are image-slider methods that we adapted to our base text-to-video (T2V) generation model to obtain video-slider baselines. For this comparison, we recruited **122 participants**. Each participant rated 20 sets of video edits, corresponding to 4 different editing instructions evaluated across 5 methods (the four video-slider-based baselines and ours).

We recruited another **90 participants** to evaluate the slider-based I2I+I2V and text-based V2V baselines. Each participant rated 12 sets of video edits, corresponding to 4 different editing instructions evaluated across 3 methods: Kontinuous Kontext [28] combined with Senorita [48], UniVideo [40], and our method.

We report the average Likert-scale ratings in Table 6, where "strongly agree" corresponds to 5 and "strongly disagree" corresponds to 1. Participants generally agreed that our method can produce progressive edits while preserving subject identity and background consistency. These results largely follow the same trends as the VLM-based evaluation, although the human ratings exhibit substantially greater variance, particularly for identity (ID) and background (BG) preservation.

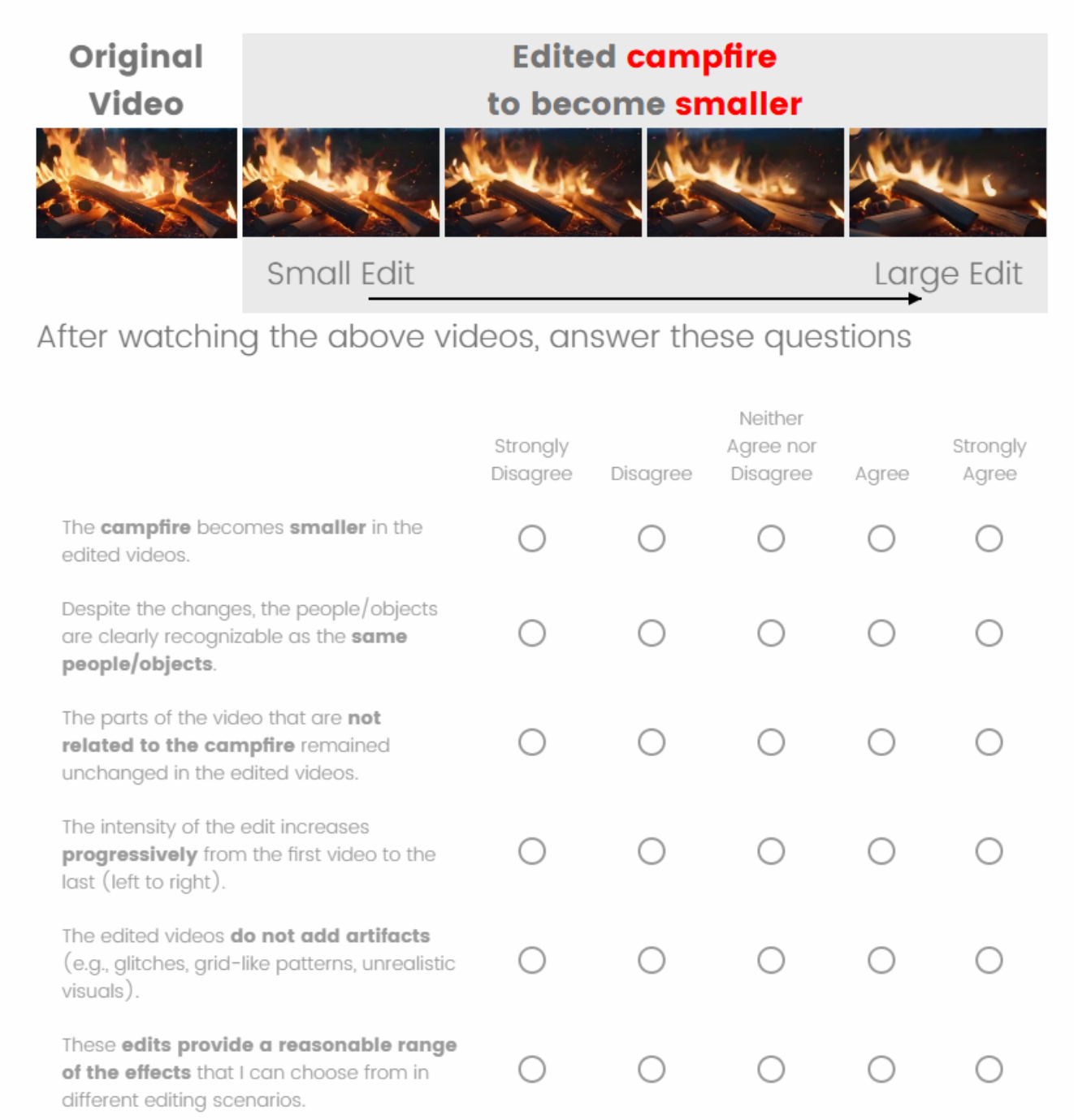

Figure 10. A screenshot example from our human evaluation study.

Participants also agreed more strongly than for other baselines that the "edits provide a reasonable range of the effects that I can choose from in different editing scenarios," as reflected in the "General Preference" column of Table 6. This suggests a clear subjective preference for our proposed editing approach.

We further visualize the Likert-scale distributions in Fig. 12. Each segment of the diverging stacked bar chart represents the percentage of participants who assigned a particular Likert rating to a given question for a given

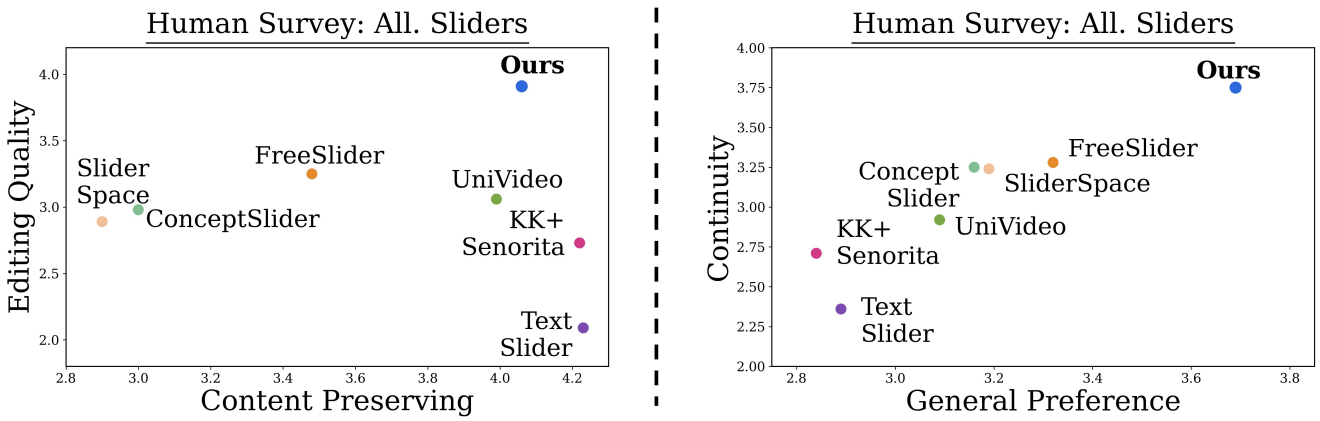

Figure 11. We compare methods across appearance sliders and motion sldiers using quadrant plots of edit quality vs. content preservation, and slider continuity vs. video general preference. The results are the combination of appearance sliders and motion sliders.

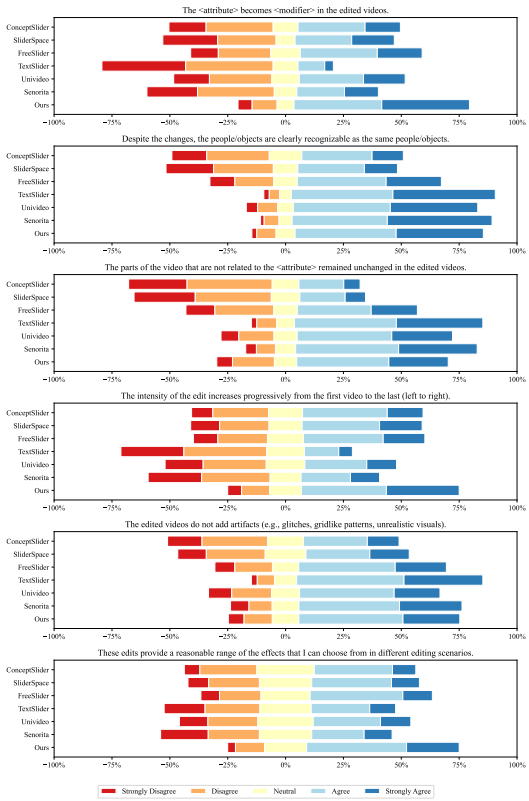

Figure 12. Distribution of Likert ratings in our human evaluation study.

baseline. Our method consistently receives more "strongly agree" and "agree" responses across most questions, all of which were phrased such that agreement indicates a positive evaluation. However, for content-preservation-related questions, Text Slider and Kontinuous Kontext + Senorita sometimes receive comparable or even higher agreement.

Table 6. User study results across 6 baselines including appearance/motion sliders.

Method	Editing Qual. ↑	Content Prev. ↑	BG. Prev. ↑	Continuity ↑	Less Artifact ↑	General Preference ↑
ConceptSlider	2.98	3.00	2.46	3.25	2.97	3.16
SliderSpace	2.89	2.90	2.51	3.24	3.12	3.19
FreeSlider	3.25	3.48	3.21	3.28	3.52	3.32
TextSlider	2.09	**4.23**	**4.05**	2.36	**4.02**	2.89
Senorita	2.73	4.22	3.95	2.71	3.71	2.84
Univideo	3.06	3.99	3.63	2.92	3.43	3.09
Ours	**3.91**	4.06	3.59	**3.75**	3.68	**3.69**

We attribute this to the fact that these methods tend to make minimal changes, which also explains their many "strongly disagree" responses on edit-quality-related questions in Fig. 12. Figure 13 compares our method with two baselines, Text Slider and Kontinuous Kontext + Senorita. Both baselines tend to make only minimal edits, which helps preserve subject identity and background appearance, but limits their ability to produce noticeable and controllable changes. Figure 14 further highlights a limitation of Kontinuous Kontext + Senorita: since it performs editing based only on the first frame, it can fail when the target object or region to be edited is absent in that frame. This limitation then propagates to the entire video, leading to weak or unsuccessful edits.

B. Details of slider-based I2I-I2V and text-based V2V

We first generate a base video from our underlying video generation model using the same prompt and same random seed. For the slider-based I2I baseline, we use Kontinuous Kontext [28], a slider-based image editing pipeline. Given an editing instruction, Kontinuous Kontext produces the corresponding edited image at different strength levels. We use the first frame of the generated video as the image to be edited. After obtaining the edited first frames, we

Figure 13. Comparison with Text Slider and Kontinuous Kontext + Senorita. Both baselines tend to make only minimal edits, which better preserves identity and background but limits edit strength and controllability.

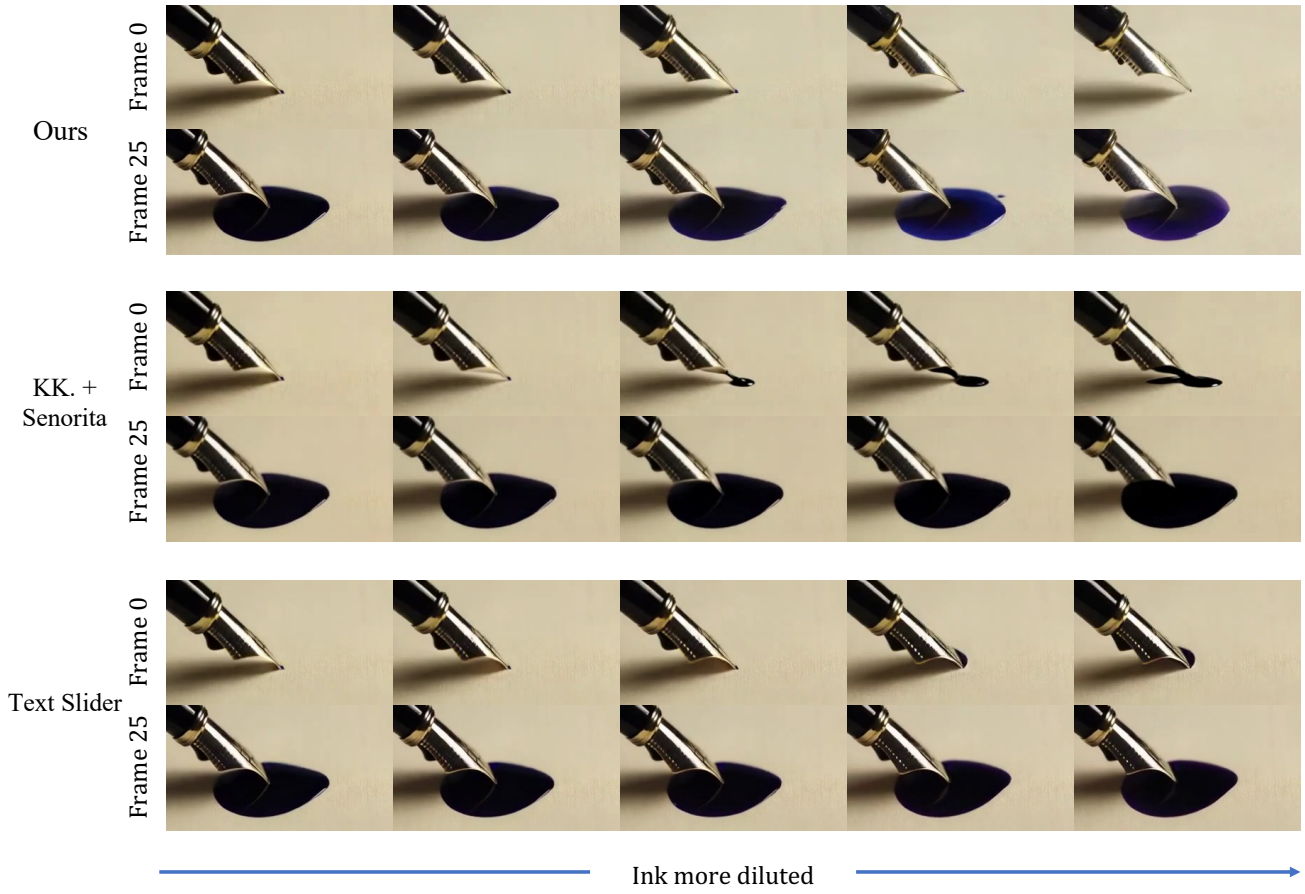

Figure 14. Failure case of Kontinuous Kontext + Senorita. Because the method edits only based on the first frame, it struggles when the target object or region is absent initially, causing the failure to propagate through the full video.

apply a first-frame propagation-based video editing model, Senorita [48]. Given the edited first frame and the original video, Senorita propagates the first-frame edit to the entire video, producing the final edited videos.

For the text-based V2V baseline, we use UniVideo [40], a text-driven video-to-video editing model. Given the original video and an editing instruction, we construct a set of modified editing instructions that explicitly encode different

Table 7. Concepts and Attributes for Evaluations

Concepts	Attributes
aurora	brighter, dimmer, larger, more purple, smaller
bubble	more likely to burst quickly, denser, larger, smaller, sparser
campfire	bluer, brighter, larger, smaller, warmer
confetti	more chaotic, more color-saturated, denser, sparser, more stable
explosion	brighter, dimmer, larger, smaller, more smoky
ink	more diluted, greener, redder, more spread out, thicker
person	more curly hair, happier, heavier, older, from walk to run
smoke	calmer, darker, thicker, thinner, more turbulent
snowflake	brighter, denser, larger particle, smaller particle, sparser
spark	brighter, more chaotic, denser, sparser, more stable
water splash	more chaotic, less droplet, more droplet, more stable, higher viscosity
speed	higher motion magnitude, lower motion magnitude

edit strengths (e.g., *slightly* → *moderately* → *much* → *extremely*). For example, if the original instruction is "make the campfire redder," the corresponding scale-specific instructions are:

- "make the campfire *slightly* redder"
- "make the campfire *moderately* redder"
- "make the campfire *much* redder"
- "make the campfire *extremely* redder"

We use these prompts as inputs to UniVideo and perform each edit independently.

C. Dataset Details

C.1. Training Dataset

Table 7 summarizes the concepts and attributes used in our evaluation. For each concept, we construct a small internal text–video paired dataset containing a few hundred training samples with minimal supervision. We first use keyword-based filtering to select text–video pairs whose text descriptions explicitly mention the target concept, followed by light manual cleaning and deduplication. This produces training pairs where the target concept is named in the prompt and usually visually centered in the corresponding video. For example, for the *campfire* concept, one training sample uses the text prompt: "A campfire in an abandoned city street at night, with urban ruins surrounding it."

For the *speed* concept, we additionally curate a few hundred green-screen training videos. The motivation for using the green-screen videos is to reduce background noise and focus the supervision more directly on foreground motion changes. Although these green-screen videos are used during training, empirically we found the learned model generalizes well to natural scenes at inference time.

C.2. Test Time Settings

At test time, for each concept–attribute pair shown in Table 7, we construct 16 base prompts that are not included in the training set. Each prompt explicitly contains the target concept. For each prompt, we generate two base videos using different random seeds. For every concept–attribute–prompt–seed combination, we then generate edited videos at scales 1–4, where scale 0 corresponds to the original video. This evaluation protocol enables us to assess controllable editing across diverse concepts, prompts, random seeds, and edit strengths.

D. Effect of Posterior Refinement

We further study the effect of posterior refinement during training. Intuitively, during reverse diffusion, the model's one-step prediction at high-noise timesteps can deviate from the desired posterior trajectory, often leading to blurry structures, weakened details, or distorted content in the reconstructed result. Such noisy and overly smoothed predictions are also likely to be out of distribution for downstream understanding models, which are typically trained on cleaner and more natural visual inputs. As a result, when these predicted clean videos are further processed by understanding models such as InternVideo2 [39] or DINOv2 [27], the extracted features and resulting supervision signals can become less reliable and more noisy.

Fig. 15 shows qualitative comparisons at several noisy timesteps during training. The second column shows the ground-truth video frames used during training. The third column shows the one-step prediction of the clean video from noisy latents at different timesteps. The fourth column shows the corresponding predicted clean video after applying posterior refinement.

Without posterior refinement, the decoded intermediate samples tend to be over-smoothed and less structurally faithful, especially for fine-grained appearance details such as flame boundaries, high-frequency textures, and local shape variations. In contrast, with posterior refinement, the intermediate predictions remain noticeably closer to the ground-truth video, yielding sharper flame structures, better local contrast, and more stable spatial layouts. This leads to more meaningful feedback from the understanding models and helps stabilize the gradients during training.

E. Lucas–Kanade Optical Flow on DINOv2 Patch Features

For dynamic attributes, the main paper defines the motion objective through a flow extractor $\mathbf{m}(\cdot)$ operating in an understanding space. In our implementation, $\mathbf{m}(\cdot)$ is instantiated as Lucas–Kanade optical flow computed on DINOv2 patch features rather than on RGB pixels. This design follows the motion-magnitude scaling formulation in the main

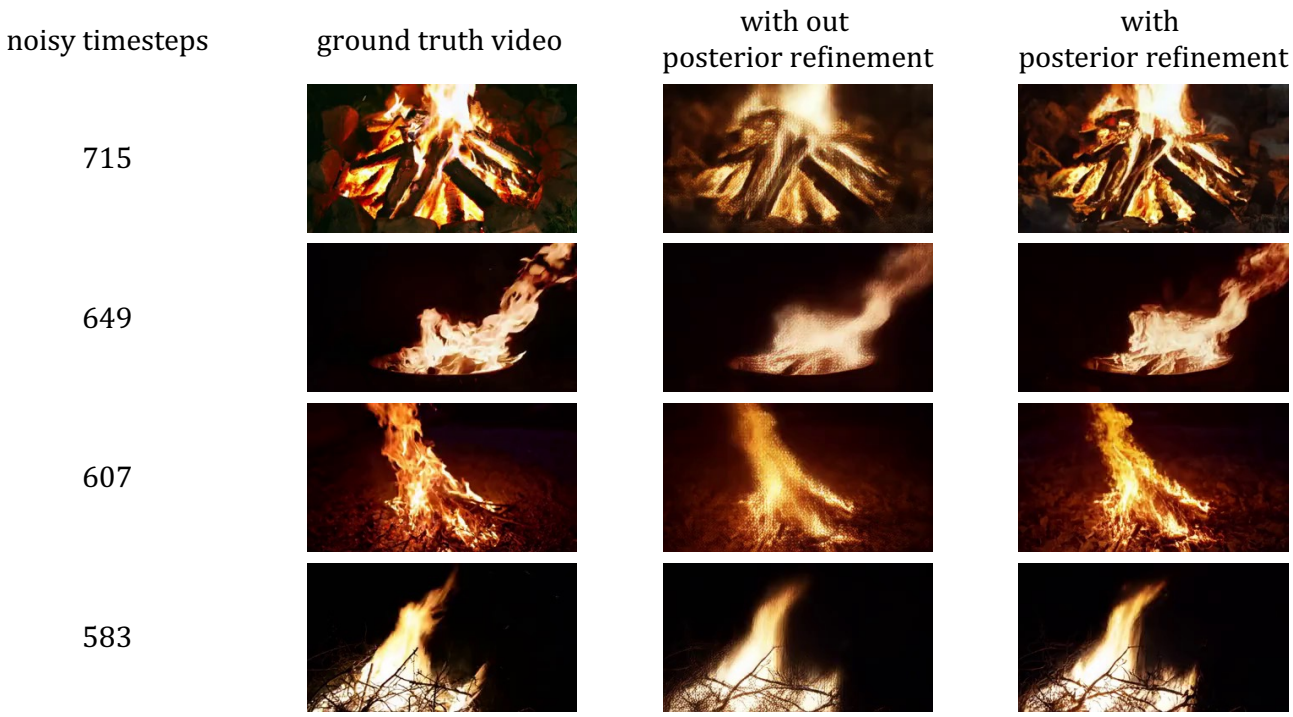

Figure 15. Qualitative comparisons at several noisy timesteps during training, the second column shows ground truth video during training, the third column shows the ons-step prediction of clean video from noisy latent at different noisy timsteps, the fourth column shows the predicted clean video with posterior refinement.

paper.

Given a predicted video $\hat{x}_0^{ref}(\Delta)$, we first sample a fixed set of keyframes and resize each frame to 224×224 before feeding them into a frozen DINOv2 encoder $\mathcal{D}$. For each frame, we extract the normalized patch tokens,

$$\mathcal{D}\Big(\hat{x}_0^{ref}(\Delta)\Big) \in \mathbb{R}^{K\times N\times D},$$

where K is the number of sampled frames, N is the number of spatial patches, and D is the patch-feature dimension. We then reshape the patch tokens into a spatial grid of size $H_p \times W_p$ with $H_pW_p = N$.

We define $\mathbf{m}\Big(\hat{x}_0^{ref}(\Delta)\Big)$ as the patch-level flow field estimated from these DINOv2 features using a multi-channel Lucas–Kanade formulation. For two consecutive frames, let $\mathbf{f}_t(i,j) \in \mathbb{R}^D$ denote the DINOv2 feature vector at patch location (i,j). We compute spatial gradients by central differences and temporal gradients by frame differences:

$$\mathbf{I}_x = \frac{\mathbf{f}_t(i,j+1) - \mathbf{f}_t(i,j-1)}{2},$$

$$\mathbf{I}_y = \frac{\mathbf{f}_t(i+1,j) - \mathbf{f}_t(i-1,j)}{2},$$

$$\mathbf{I}_t = \mathbf{f}_{t+1}(i,j) - \mathbf{f}_t(i,j).$$

Under the Lucas–Kanade assumption, the local flow (u, v) is obtained by solving a least-squares system jointly over all feature channels. In practice, we first form the structure-tensor terms

$$I_{xx} = \sum_d I_x^{(d)} I_x^{(d)},$$

$$I_{yy} = \sum_d I_y^{(d)} I_y^{(d)},$$

$$I_{xy} = \sum_d I_x^{(d)} I_y^{(d)},$$

$$I_{xt} = \sum_d I_x^{(d)} I_t^{(d)}, \quad I_{yt} = \sum_d I_y^{(d)} I_t^{(d)},$$

average them within a local spatial window, and then solve the resulting 2×2 linear system independently at each patch location. This produces a patch-level flow field

$$\mathbf{m}\Big(\hat{x}_0^{ref}(\Delta)\Big) \in \mathbb{R}^{(K-1)\times N\times 2}.$$

Our implementation uses replicate padding for boundary handling, central differences for spatial gradients, a small square averaging window for local coherence, and Cramer's rule to solve the 2×2 system.

To enforce motion magnitude scaling, we follow the stop-gradient design in the main paper and construct the target flow by scaling a detached reference flow:

$$\mathbf{m}_{\text{ref}} = \gamma \cdot \Big[\mathbf{m}\Big(\hat{x}_0^{ref}(\Delta)\Big) .\text{sg}()\Big],$$

where γ is the motion scaling factor. The dynamic loss is then

$$\mathcal{L}_{\text{dyn}} = \Big\|\mathbf{m}\Big(\hat{x}_0^{ref}(\Delta)\Big) - \gamma \cdot \Big[\mathbf{m}\Big(\hat{x}_0^{ref}(\Delta)\Big) .\text{sg}()\Big]\Big\|_2^2.$$

In implementation, the detached branch is computed without gradient, while the flow extracted from the current prediction remains differentiable with respect to the generated video frames.

F. Implementation Details on Wan 2.1

We use Wan 2.1 T2V 1.3B [37] as the base video generation model. Consistent with the formulation in the main paper, TokenDial is implemented by injecting additive token offsets into intermediate video patch tokens of the pretrained video DiT. On Wan 2.1, we realize this intervention inside each DiT block by adding the learned offset to the self-attention residual branch. Concretely, after the hidden tokens are normalized and modulated by the timestep-conditioned adaptive parameters, they are passed through the self-attention layer; the predicted token offset is then added to the self-attention output before the gated residual update. The subsequent cross-attention and feed-forward branches are kept unchanged.

Our choice of the self-attention residual branch is motivated by both architectural considerations and empirical ablations. We experimented with injecting the token offset at several alternative locations, including before and after the modulation layers, in the cross-attention branch, and after each full DiT block. Among these choices, injecting into the self-attention residual branch consistently provided the best trade-off between editability and preservation. In particular, this location better preserves the original video structure while still enabling meaningful attribute-level edits. Intuitively, the self-attention branch primarily operates on the model's latent visual tokens and thus offers a natural intervention point for modifying internal visual semantics without excessively disturbing the text-conditioning pathway or the overall block computation. By contrast, injecting into cross-attention or after the full block tends to produce less stable edits or weaker structural preservation.

This implementation preserves the core design of our method, namely, learning attribute-specific additive offsets in the intermediate token space rather than modifying backbone weights, while adapting it to the architectural structure of Wan 2.1.

G. VLM Prompt for Slider Evaluation

We use the following prompt template for VLM-based slider evaluation.

```
You are a meticulous video editing quality evaluator.

Your task is to assess a VIDEO EDIT SLIDER by analyzing 5 images sampled from the SAME timestamp,
corresponding to 5 increasing slider scales.

Scale definition:
- Scale 1: Original video frame (before editing)
- Scale 2-5: Edited frames with progressively stronger edit strength

Concept: "{concept}"
Target attribute direction: "{attribute}"
Goal: Make the concept become more "{attribute}" progressively from Scale 2 to Scale 5.

Instructions:
Analyze the 5-scale sequence as a whole and evaluate how well the editing slider satisfies the editing
    goal while preserving visual consistency.
- All 5 images depict the SAME moment in time.
- The main subject identity and background should remain the SAME across all scales despite the edit.
- Stronger edits NEVER justify replacing the subject or changing the background.

Score the ENTIRE 5-scale sequence on 4 criteria, each integer in [0,5] where higher is better.
Provide a short justification for each score.

You will evaluate the slider across FOUR criteria.
For each criterion, provide a score from 0 (worst) to 5 (best) and a brief justification.


1. Prompt Following (Score: 0-5)
Question:
Does the slider edit direction correctly and consistently make the SAME "{concept}"
appear more "{attribute}" from Scale 2 to Scale 5?

Scoring Guide:
- 5: The edit perfectly follows the prompt, with Scale 5 being the strongest and most aligned.
- 4: The edit mostly follows the prompt with minor weaknesses.
- 3: The edit partially follows the prompt but is ambiguous or inconsistent.
- 2: The edit weakly reflects the prompt.
- 1: The edit barely relates to the prompt.
- 0: The prompt is ignored or contradicted.

2. Identity Preserving (Score: 0-5)
Question:
Does the main subject remain the SAME identity and category across all scales?

If the subject is replaced, changes category, or becomes a different person/object at ANY scale,
this is considered a severe violation and will be heavily penalized.

Scoring Guide:
- 5: Identity is perfectly preserved across all scales.
- 4: Very minor visual changes but clearly the same subject.
- 3: Noticeable drift, but identity is still mostly recognizable.
- 2: Major identity inconsistency or partial replacement.
- 1: Severe identity change.
- 0: The subject is completely replaced or unrecognizable.

3. Background Consistency (Score: 0-5)
Question:
Have the regions that should NOT be edited (background, scene layout, camera viewpoint)
remained stable across all scales?

Scoring Guide:
- 5: Background is perfectly preserved and stable.
- 4: Minor, subtle background changes.
```

```
- 3: Noticeable but non-catastrophic background drift.
- 2: Significant background changes or redraw.
- 1: Severe background inconsistency.
- 0: Background is completely altered.

4. Progressive Intensity (Score: 0-5)
Question:
Does the edit strength increase monotonically and smoothly from Scale 1 to Scale 5?

Scoring Guide:
- 5: Smooth, monotonic increase with clear ordering from weak to strong.
- 4: Mostly monotonic with minor irregularities.
- 3: Inconsistent progression.
- 2: Weak or unclear progression.
- 1: Reversed or chaotic progression.
- 0: No meaningful progression.

Return STRICT JSON ONLY (no markdown) with this schema:
{
  "prompt_following": {"score": <int 0-5>, "reason": <string>},
  "id_preserving": {"score": <int 0-5>, "reason": <string>},
  "background_consistency": {"score": <int 0-5>, "reason": <string>},
  "progressive_intensity": {"score": <int 0-5>, "reason": <string>}
}
```

H. More Results

More results for appearance sliders and motion sliders can be found in the attached video presentation, and the index.html file inside the project page folder.

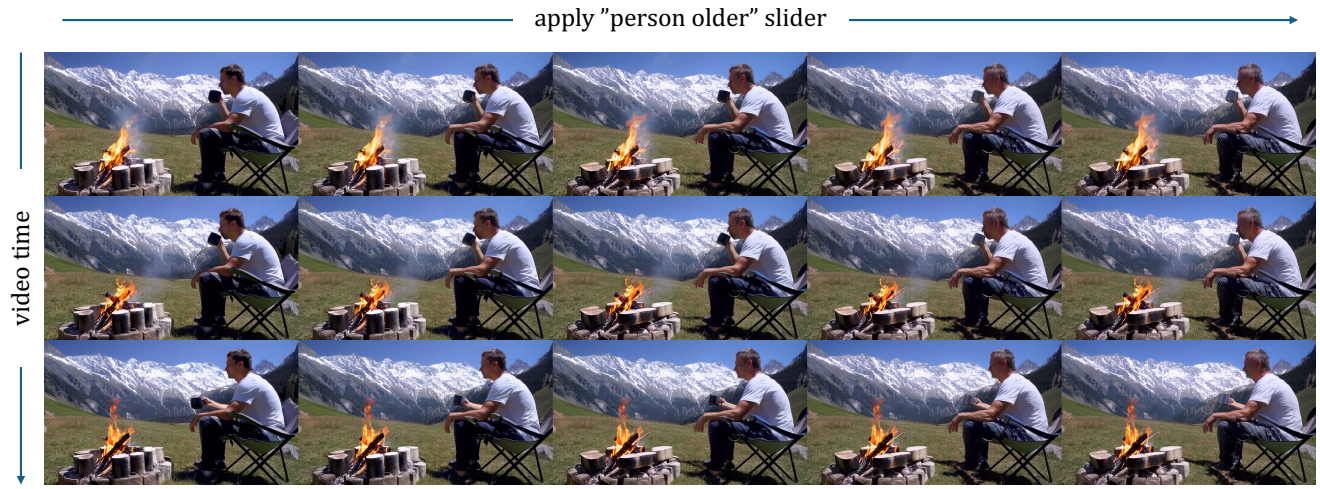

Figure 16. Additional results for appearance slider.

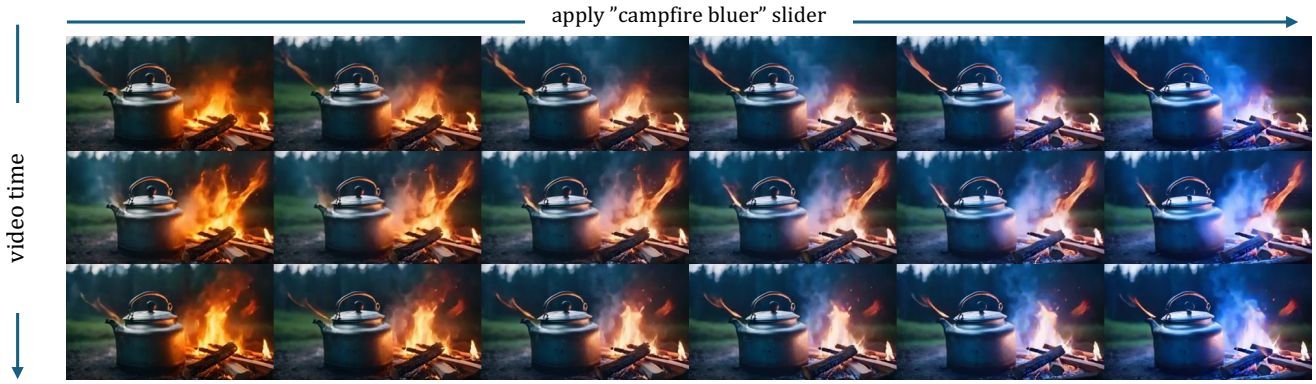

Figure 17. Additional results for appearance slider.

Figure 18. Additional results for appearance slider.

Figure 19. Additional results for appearance slider.