MatchDiffusion: Training-free Generation of Match-Cuts

Alejandro Pardo*[1] Fabio Pizzati*[2,3] Tong Zhang[1] Alexander Pondaven[3]
Philip Torr[3] Juan Camilo Perez[1†] Bernard Ghanem[1]

[1]KAUST [2]MBZUAI [3]University of Oxford

*Equal contribution. †Work done while at KAUST, now at Meta.

https://matchdiffusion.github.io

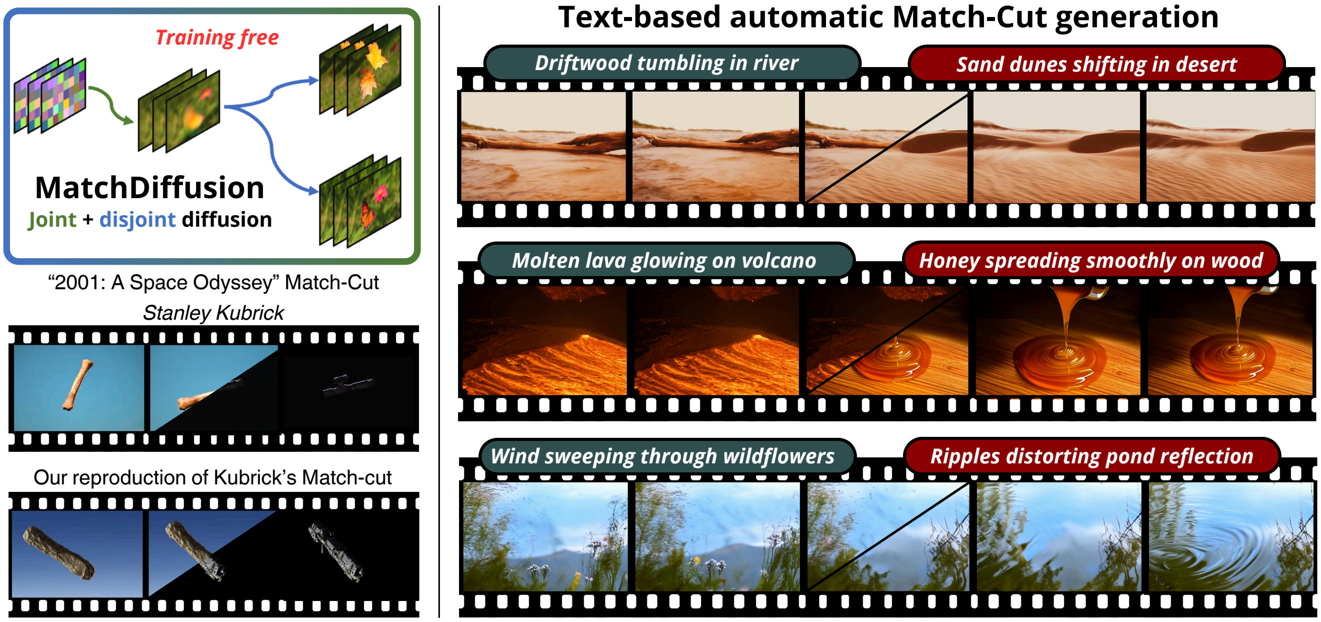

Training free
MatchDiffusion
Joint + disjoint diffusion
"2001: A Space Odyssey" Match-Cut
Stanley Kubrick
Our reproduction of Kubrick's Match-cut
Text-based automatic Match-Cut generation
Driftwood tumbling in river
Sand dunes shifting in desert
Molten lava glowing on volcano
Honey spreading smoothly on wood
Wind sweeping through wildflowers
Ripples distorting pond reflection

Figure 1. **Automatic match-cut generation with MatchDiffusion.** In the history of cinema, there is prevalent use of match-cut transitions, *i.e*. semantic shifts in the content of two scenes that share the same structure, as exemplified by Stanley Kubrick's iconic transition from a bone to a spaceship (bottom left). However, obtaining visually appealing match-cuts requires sophisticated planning and multiple shots, due to the complexity of the transition. Our proposed MatchDiffusion approach is able to automatically generate match-cuts following textual prompts (right), thanks to a training-free inference technique composed of Joint and Disjoint Diffusion mechanisms (top left).

Abstract

Match-cuts are powerful cinematic tools that create seamless transitions between scenes, delivering strong visual and metaphorical connections. However, crafting match-cuts is a challenging, resource-intensive process requiring deliberate artistic planning. In MatchDiffusion, we present the first training-free method for match-cut generation using text-to-video diffusion models. MatchDiffusion leverages a key property of diffusion models: early denoising steps define the scene's broad structure, while later steps add details. Guided by this insight, MatchDiffusion employs "Joint Diffusion" to initialize generation for two prompts from shared noise, aligning structure and motion. It then applies "Disjoint Diffusion", allowing the videos to diverge and introduce unique details. This approach produces visually coherent videos suited for match-cuts. User studies and metrics demonstrate MatchDiffusion's effectiveness and potential to democratize match-cut creation. Visit our website for video results. Our code is open source.

1. Introduction

> *"The art challenges the technology,*
> *and the technology inspires the art."*
>
> – John Lasseter

Cinematic transitions are powerful storytelling tools that evoke emotions, suggest the passage of time, or visually connect themes [28]. Among transitions, *match-cuts* are defined as “Two successive shots joined to create a strong similarity of compositional elements” [34]. They are particularly effective in bridging two scenes with strikingly different content but similar composition, creating a strong sense of visual continuity. For instance, in Kubrick’s “2001: A Space Odyssey”, a match-cut is used to transition from a bone thrown by an ape into a satellite orbiting Earth—conveying Humanity’s evolutionary leap, from primitive tools to space technology, without a single word.

Despite their visual elegance and narrative power, match-cuts are notoriously difficult to create. They require careful planning and precise visual alignment, often shaping the entire production process to ensure a cohesive transition [1, 29, 33, 42, 44]. This complexity limits match-cuts to experienced filmmakers with substantial resources, making them rare cinematic gems. Our aim is to democratize this powerful tool by providing a simple method that allows creators of various skill levels to experiment with match-cuts, helping amateurs and experienced filmmakers to quickly iterate and refine ideas before full-scale production.

While many video editing tasks—such as scene interpolation [11], video morphing [27], and multi-text generation [36]—create continuity by generating intermediate frames, gradually deforming visuals, or extending narratives across prompts, match cuts operate differently. Rather than blending scenes, they establish a direct visual connection between *two distinct yet compositionally aligned shots*. Leveraging this principle, we formulate match-cut generation as synthesizing a *pair of videos* that share structural coherence while differing in semantics, further post-processed into a match-cut. To achieve this, we exploit an empirical property of text-to-video diffusion models, extending prior diffusion analyses [8, 23, 35]. We introduce MatchDiffusion, a *training-free* approach that generates match-cuts from two prompts by guiding the diffusion process.

Our method first performs “Joint Diffusion”, by initializing the synthesis for both prompts from a single noise sample and then guiding both along a common denoising path for the first denoising steps. This process translates into a cohesive layout and structure being shared between the two videos. After this stage, we then perform “Disjoint Diffusion”, where we allow the videos’ diffusion paths to diverge, as guided by their corresponding prompts. With these processes, MatchDiffusion generates videos that independently exhibit unique content while jointly displaying visual coherence established in the early stages—resulting in distinct yet harmonized scenes suitable for a match-cut. Please refer to Fig. 1 for an overview of our approach.

To evaluate our diffusion-based approach to synthesizing match-cuts, we implement intuitive baselines using existing methods (*e.g.* [27, 50, 54]). We selected each of these methods for its potential to effectively assess aspects of the generation of match-cuts. Alongside these baselines, we propose metrics to quantify match-cut quality, and allow comparing synthesis methods. Together, these elements establish an evaluation framework that demonstrates our method’s effectiveness and adaptability.

In summary, our contributions are three-fold: **(i)** We formalize the task of creating match-cuts as synthesizing video pairs that are structurally coherent yet semantically divergent. **(ii)** We introduce MatchDiffusion, a training-free method that leverages pre-trained diffusion models to automate the generation of match-cuts. **(iii)** We implement robust baselines and propose metrics for evaluating match-cut quality, establishing a benchmark for synthesis methods.

2. Related works

Noise manipulation in diffusion models Early works showed that manipulating noise in diffusion models can yield desirable outputs. In images, fusing noise estimates from multiple prompts enables composition [5, 57], while Visual Anagrams [13] and Factorized Diffusion [12] reveal how summing noise can produce optical illusions. SyncDiffusion [20] applies similar ideas to generate panoramic images. SyncTweedies [18] unifies such strategies, including MultiDiffusion [5], under a general framework. In videos, noise is often fused across all iterations to ensure long-term consistency [36, 45, 46] or enable scene interpolation [11], with emerging properties over time also explored [7]. Crucially, these methods combine noise across *all* steps to produce a single scene. In contrast, we formulate match cuts as a distinct problem, *exploiting diffusion’s emerging properties* in a novel way to generate coherent video pairs.

Match-cut synthesis. Cutting in video editing has been widely explored. Some focus on detecting cut points in untrimmed videos using audio-visual cues [30], audio-beat alignment [32], or transitions for dialogue scenes [19], without differentiating types of transitions. Shen *et al.* [40] propose smooth transitions such as fades and panes, excluding straight cuts. Pardo *et al.* [31] offer a dataset for straight-cut classification, with match-cuts as one category, though underrepresented. Recently, retrieval-based approaches addressed match-cut creation: one curating candidates via audio-visual features [9], the other focusing on audio-based match-cuts [10]. These studies tackle match-cut synthesis through retrieval, whereas we propose a generative approach to synthesize video pairs that form a match-cut.

Multi-scene video generation. Recent work has explored multi-shot video generation. Several works [22, 25, 26] generate multi-scene layouts from scripts derived by large language models (LLMs), while others [14, 51, 56] focus on ensuring temporal coherence and reducing hallucinations between frames. TALC [4] improves temporal alignment

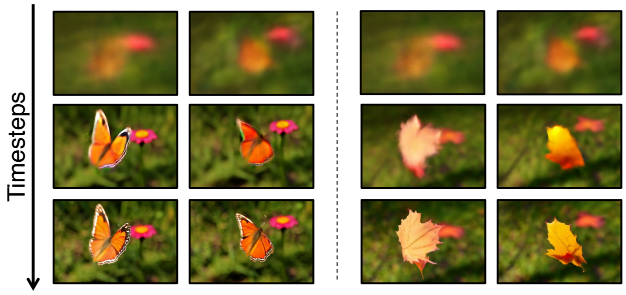


Figure 2. **Feature emergence during denoising.** While the first iterations (top) yield ambiguous outputs displaying colors and basic structure, further iterations inject semantics (middle), until the final output is generated (bottom).

with aligned captions, while Contrastive-Sequential Diffusion [37] enhances visual coherence in multi-scene videos. MiNT [49] allows for cut synthesis, but maintains high semantic consistency. Unlike these, which focus on ensuring consistency within a single narrative, our method prioritizes structural coherence across two *distinct* prompts, optimizing for structure and motion consistency suited for match-cut transitions rather than narrative flow.

3. MatchDiffusion

Given two prompts (ρ', ρ'') describing different scenes, our goal is to generate a pair of videos (x', x'') that align with their respective prompts while remaining visually cohesive for match-cut transitions. Each video is generated separately, allowing them to be seamlessly post-processed into a match-cut. For example, this can be done by joining the first half of x' with the second half of x''. This setup enables multiple cut styles from a single video pair. On our website, we showcase three such styles: *straight cuts*, *alpha blending*, and *flickering*. While these are just a few editorial options, the synchronized outputs open the door to many other creative possibilities. We rely on a key property of diffusion models to achieve paired video synthesis: as highlighted in prior works [8, 23, 35] and illustrated in Figure 2, diffusion models establish broad structural patterns in the early denoising stages, while finer details and prompt-specific textures emerge later. By leveraging this progression, we design MatchDiffusion, a two-stage training-free pipeline tailored for match-cut generation. MatchDiffusion comprises: (1) Joint Diffusion (Section 3.2), where we set up a shared visual structure based on both prompts, followed by (2) Disjoint Diffusion (Section 3.3), where each video independently develops the semantics corresponding to its prompt. In the following sections, we introduce preliminaries, then go into detail into each stage of MatchDiffusion, elaborating on how the joint and disjoint diffusion stages provide the balance needed for match-cut generation.

3.1. Preliminaries

We first introduce the working mechanism of diffusion models for text-to-video (T2V) synthesis. T2V models operate by iteratively denoising Gaussian noise, with the goal of producing a fully denoised video that aligns with a conditioning textual prompt. Recent methods [53] execute this process in a latent space established by a pretrained autoencoder, mitigating computational costs [39]. The autoencoder comprises an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The latent space of this autoencoder is then iteratively denoised by a noise estimation network ϵ_θ over T steps, starting from sampled Gaussian noise $z_T \sim \mathcal{N}(0, I)$. We denote the latent video representation at the t-th iteration as z_t, where $t \in \{0, ..., T\}$. That is, the network ϵ_θ predicts the noise ϵ_t for z_t. The network's prediction is conditioned on both the input textual prompt ρ and the timestep t: $\epsilon_t = \epsilon_\theta(z_t, \rho, t)$.

This noise prediction is then used to update the noisy latent representation, following scheduling strategies such as DDPM [16] or DDIM [41]. Namely, at step t, the noisy representation z_t is denoised into $z_0^{(t)}$ by combining the estimated noise with the latent representation: $z_0^{(t)} = z_t - \gamma_t \epsilon_t$, where γ_t is a scaling factor function of t. Then, another Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ sample is used to noise $z_0^{(t)}$ again, following a noise schedule whose intensity decreases over timesteps. Formally: $z_{t-1} = \eta_t z_0^{(t)} + \sigma_t \epsilon$, where η_t and σ_t regulate the noise intensity, which decreases with increasing t [16, 41]. After T timesteps, z_0 is decoded via $x = \mathcal{D}(z_0)$ into the output video x.

For creating a match-cut, we *generate two videos simultaneously* by breaking the diffusion process into two stages: a *joint* stage where the latent representation of the videos is shared, and a *disjoint* one where representations are allowed to diverge. We now elaborate on each.

3.2. Joint Diffusion

The first stage of MatchDiffusion is Joint Diffusion. During this stage, we simultaneously generate both videos by forcing the synthesis to incorporate *both input prompts* for the first K denoising iterations, where $K \in \{0, ..., T\}$. After these K iterations, the result is a single latent displaying an abstract structure that broadly satisfies *both* prompts. Our intuition behind this design builds on previous work on hybrid images [6, 12, 13], showing that the diffusion process can be manipulated to produce images displaying different scenes depending on viewing conditions. However, our scenario is unique, since we require each output, x' and x'', to clearly and independently comply with its own prompt, sharing only selected appearance-related traits. As illustrated in Figure 2, the intermediate denoising outputs $z_0^{(t)}$ reveal motion patterns and the scene layout—the essential elements for match-cuts—emerge in early stages, while later refinement steps focus on details related to semantic

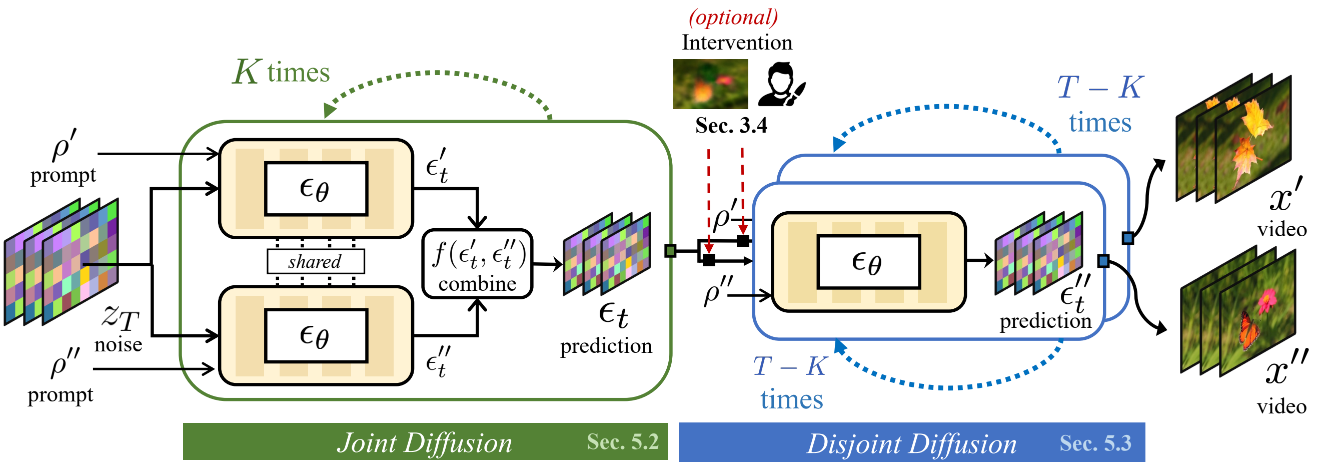


Figure 3. **MatchDiffusion.** We formulate the task of creating match-cuts as generating a pair of videos sharing a general appearance while having different in semantics. A portion of the frames of these videos can then be combined to enable match-cut transitions. To generate these videos, MatchDiffusion first performs a Joint Diffusion process for K steps (left) by combining the noise predictions from the two prompts via a function f. Then, a Disjoint Diffusion process is executed to obtain the final outputs x' and x'', *i.e.* denoising separately for the remaining $T-K$ iterations with one prompt per path. Optionally, MatchDiffusion also supports manual user intervention by allowing the integration of generated video tone and structural edits.

content. As shown in Figure 3 (left), for the first K iterations, we combine noise predictions from each prompt using a function f, ensuring shared foundational characteristics early in synthesis. The joint diffusion process is defined by modifying noise estimate to:

$$\epsilon_t = f(\epsilon_\theta(z_t, \rho', t), \epsilon_\theta(z_t, \rho'', t)), \tag{1}$$

while keeping the computation of z_{t-1} unchanged, as described in Section 3.1. Although this formulation supports different expressions for f, we choose it to simply be the averaging function, *i.e.* $f(a, b) = {}^{(a+b)}/_{2}$.

3.3. Disjoint Diffusion

After K iterations of Joint Diffusion, we obtain a noisy latent z_{T-K} encoding characteristics that are desirable to preserve in both x' and x''. This second stage of Disjoint Diffusion allows the remaining $T-K$ steps of the diffusion process to start from this latent but depart from the shared path to introduce the characteristics that are specific to the individual prompts. In particular, Disjoint Diffusion starts from z_{T-K} and finishes denoising via $T-K$ evaluations of ϵ_θ, conditioned on one prompt at a time. As such, Disjoint Diffusion produces separate noise predictions ϵ'_t and ϵ''_t, as shown in Figure 3 (right). This procedure ensures that the emergence of semantics and details specific to each prompt occurs while maintaining the structure encoded in the initial K steps. For $t \in \{0, ..., T-K\}$, this becomes:

$$\epsilon'_t = \epsilon_\theta(z'_t, \rho', t), \quad \epsilon''_t = \epsilon_\theta(z''_t, \rho'', t). \tag{2}$$

When $t = T-K$, both z'_t and z''_t are set to z_{T-K}. After Disjoint Diffusion, we obtain two videos, $x' = \mathcal{D}(z'_0)$ and $x'' = \mathcal{D}(z''_0)$, which can be combined into a match-cut.

One might assume that results of MatchDiffusion resemble those of video-to-video translation based on SDEdit [27], which perform prompt-based editing by injecting noise into an existing video x_{init} from step K onward. However, our approach is fundamentally different, as we jointly synthesize the two scenes, rather than modifying an initial video. That is, MatchDiffusion generates outputs that *satisfy both prompts from scratch*, effectively narrowing the range of possible appearances to those that align with the shared structure and characteristics specified by both prompts. This process enables the synthesis of match-cuts for semantically uncorrelated scenes, as shown in Fig. 6, where the video-to-video translation approach fails.

User intervention. To allow for iterative user editing, we propose a human-in-the-loop strategy for a finer customization of the generated videos. Namely, a user may wish to depart from the strict color adherence of the match-cut to better align with the tone of a preceding sequence, or to modify the background. While this could be achieved with post-processing, we propose a more natural mechanism that integrates user interventions directly *into* the diffusion process.

We define τ as a generic user-driven modification, which may be automatic (*e.g.*, a color look-up table) or manual (*e.g.*, adding scene elements). We incorporate τ in the denoised video at the start of a disjoint diffusion path, *e.g.* $x_0^{(K)} = \mathcal{D}(z_0^{(K)})$, as shown in Fig. 4. By doing so, we

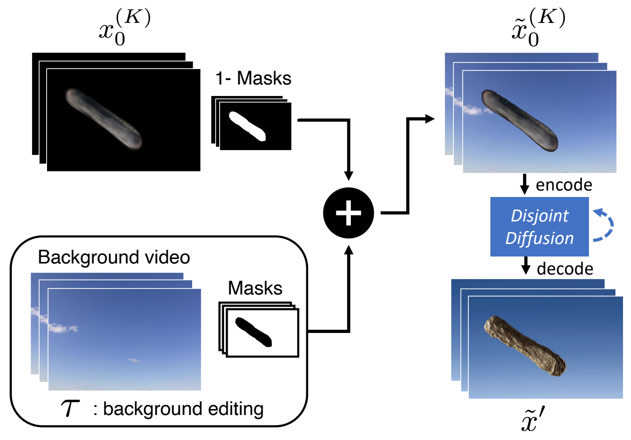


Figure 4. **User intervention.** For reproducing the match-cut in the teaser, we apply a background mask to the denoised output generated by joint diffusion. After the remaining denoising iterations, the output is refined to integrate the new background.

obtain an updated video *i.e.* $\tilde{x}_0^{(K)} = \tau(x_0^{(K)})$. We then encode this video into its corresponding $\tilde{z}_0^{(K)}$ and proceed with disjoint diffusion. Hence, we integrate τ seamlessly into the synthesized video by leveraging the diffusion process itself to achieving realistic modifications. Although some [52] have proposed similar intervention mechanisms for some τ such as background editing, our novel integration with Disjoint Diffusion allows us to preserve a coherent match-cut. Importantly, since the diffusion process continues for $T-K$ steps after τ's application, even modifications that would otherwise compromise scene realism in post-processing will be inherently refined, as shown in Fig. 4.

4. Experiments

We provide our experimental setup in Section 4.1, then show results of the match-cuts generated by MatchDiffusion in Section 4.2. Later, we compare against baselines, using qualitative and quantitative evaluations as well as user studies, in Section 4.3. We further report results with potential user interventions in Section 4.4. We conclude with a sensitivity analysis of MatchDiffusion to K in Section 4.5.

4.1. Setup

MatchDiffusion settings. For the backbone of MatchDiffusion, we choose the open-source text-to-video (T2V) diffusion model CogVideoX-5B [53]. For sampling, we use a DDIM scheduler [41] with $T = 50$ steps. For baselines and ours, we generate videos with 40 frames, and form a match-cut by concatenating the first 20 frames of x' with the last 20 of x''. We tune K for each pair of prompts. Our method's computational cost is the same as the backbone (i.e. CogVideoX-5B) since it does not add any extra computational cost. We report results on 50 prompt pairs.

Baselines. To the best of our knowledge, we are the first to synthesize match-cuts from scratch. Hence, the definition of suitable baselines is challenging. We define here three strong baselines in our best efforts to define different strategies for training-free match-cut synthesis:
Video-to-video. We define a video-to-video (V2V) translation baseline, and note that these approaches are designed for structural consistency. We first use ρ' to generate x' with the T2V version of CogVideoX-5B. Then, we use the V2V version of the same model (based on SDEdit [27]) to inject noise at step K in x', and denoise using ρ'', obtaining x''.
Motion Transfer. Recent literature has highlighted the possibility of conditioning the generation of new videos with the motion of an existing video. These motion transfer approaches allow for disentangling the motion from the reference scene content. Compared to V2V, this approach increases the flexibility in the outputs, allowing to significantly depart from the appearance of the reference video. We use a T2V model to generate x' from ρ', then we use either SMM [54] or MOFT [50] to synthesize a new video with ρ'' as input, and x' as guidance. For a fair comparison, we reimplemented SMM and MOFT on top of CogVideoX-5B. Hence, *all our baselines use the same backbone*. In supplementary, we include extra analysis on the suitability of these baselines for match-cuts, along with experiments using alternative backbones and V2V methods.

Metrics. The evaluation of a match-cut is highly subjective. However, we propose different metrics to quantify desirable aspects of a match-cut. First, we use a frame-wise CLIPScore [15] to assess prompt adherence of the generated video. We average the CLIPScore of x' and ρ', and x'' and ρ'' for each frame. This ensures that each video respects its prompt. To evaluate motion agreement between x' and x'', we use the Motion Consistency metric of SMM [54], evaluating the motion consistency of tracklets extracted by a pre-trained tracking model [17]. Finally, we use LPIPS [55] to quantify frame-wise perceptual similarity across x' and x''. Intuitively, a low LPIPS should indicate structurally-consistent outputs, *i.e.* suitable for match-cuts. In Supp., we measure the video quality of our method's results and show that it does not yield to any quality degradation.

4.2. Results

We report outputs of MatchDiffusion in Figs. 1, 5, and 6. In Fig. 5, we show a variety of match-cuts generated by our method, highlighting its ability to connect diverse concepts across different scenes. In the first two rows, MatchDiffusion demonstrates capacity to bridge unrelated scenes through background elements. For example, in the lighthouse scene, the beam of light seamlessly transitions into the fog of the adjacent scene, creating a cohesive visual connection. The third row illustrates a color-based match: transitioning from a spice market to a painter's palette by

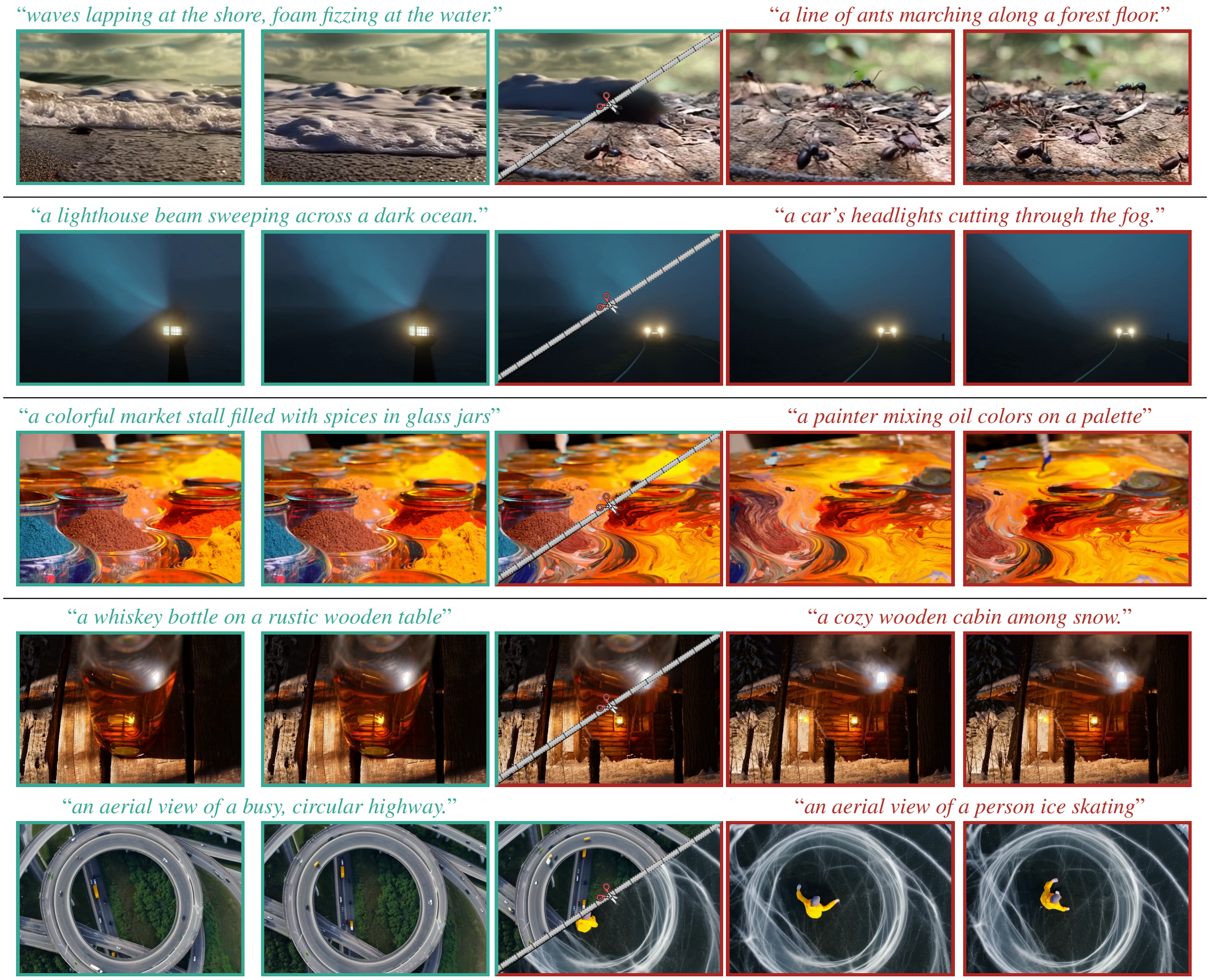

"waves lapping at the shore, foam fizzing at the water."
"a line of ants marching along a forest floor."
"a lighthouse beam sweeping across a dark ocean."
"a car's headlights cutting through the fog."
"a colorful market stall filled with spices in glass jars"
"a painter mixing oil colors on a palette"
"a whiskey bottle on a rustic wooden table"
"a cozy wooden cabin among snow."
"an aerial view of a busy, circular highway."
"an aerial view of a person ice skating"

Figure 5. **Generated match-cuts.** MatchDiffusion can automatically synthesize match-cuts based on the prompts in green and red. Note how the cuts enjoy highly consistent appearance while preserving each prompt's semantics. Please see the website for more samples.

aligning the colors in each scene. The last two rows highlight structural alignment across scenes. In the fourth row, the shape of a bottle transitions into a wooden cabin, exploiting how the liquid's color mirrors the hues of the cabin. The final row connects a highway with an ice-skating scene, aligning the circular highway shape with the ice ring's structure. These examples demonstrate the ability of MatchDiffusion to generate creative match-cuts that would otherwise be challenging to envision.

4.3. Comparison with baselines

Qualitative comparison. Fig. 6 displays frames before and after the transition for three different prompts, comparing MatchDiffusion with our proposed baselines. This figure illustrates how each approach handles various cases of match-cuts. As seen in the first column, V2V tends to produce similar-looking scenes across prompts. This result is expected, as these methods are primarily designed to translate features within scenes that already share visual similarities (*e.g.*, changing the season from summer to winter). When faced with highly dissimilar prompts, V2V typically alters minor aspects of the scene, which fall short of achieving the strong semantic shifts needed for a high-quality match-cut. For example, in the first row, the burning parchment merely becomes more rounded in the subsequent frame. Instead, motion transfer methods, such as SMM and MOFT, yield results aligned with the prompts, preserving movement across frames. However, in the same example, we observe that SMM and MOFT depart significantly from the appearance of the original image, preventing the structural alignment present in match-cuts. Finally, MatchDiffusion achieves smoother and cohesive transitions by aligning

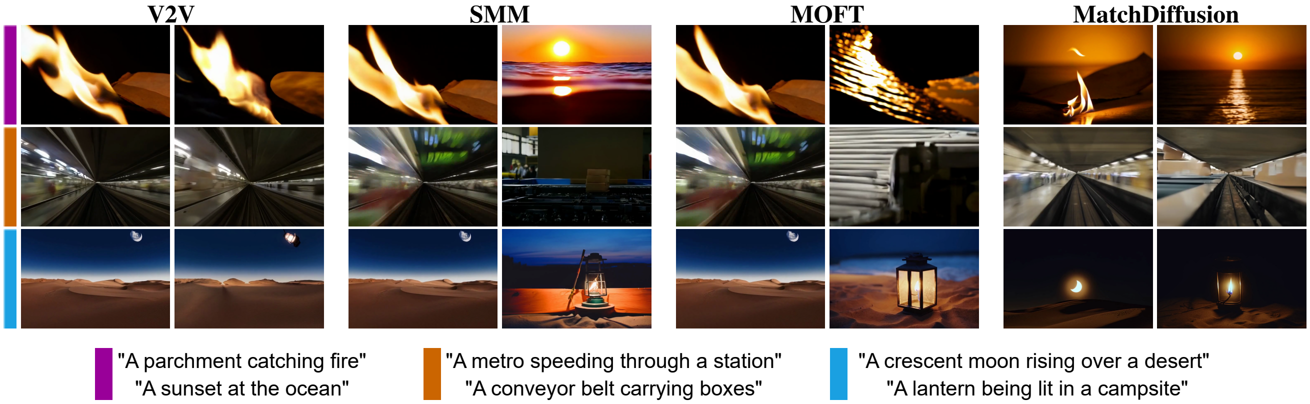

V2V
SMM
MOFT
MatchDiffusion
"A parchment catching fire"
"A sunset at the ocean"
"A metro speeding through a station"
"A conveyor belt carrying boxes"
"A crescent moon rising over a desert"
"A lantern being lit in a campsite"

Figure 6. **Qualitative comparison with baselines.** Overall, we notice that V2V does not allow for drastic modifications of the scene in presence of prompts with strong semantic differences (*e.g.*, first row). On the contrary, motion transfer baselines (SMM and MOFT) depart significantly from the content of the scene, prohibiting for a visually-appealing match-cut. Only MatchDiffusion achieves a satisfying balance between semantic changes and prompt consistency.

Method	**CLIPScore ↑**	**Motion ↑**	**LPIPS ↓**
T2V (Lower Bound)	0.33	0.40	0.74
V2V	0.31	0.67	**0.31**
SMM	**0.34**	0.64	0.74
MOFT	0.33	0.66	0.56
MatchDiffusion	0.34	**0.70**	0.32

Table 1. **Metrics comparison.** Aligned with qualitative results (Fig. 6), we report that V2V is mostly impacted in CLIPScore, due to many translations not being able to follow the prompts. On the other hand, SMM and MOFT excessively modify the scene, resulting in a high LPIPS. Only MatchDiffusion allows for high performance in all metrics. Best results are **boldfaced**, second best are underlined. Red cells show the worst performing scores. Our method (gray) strikes the best balance among all.

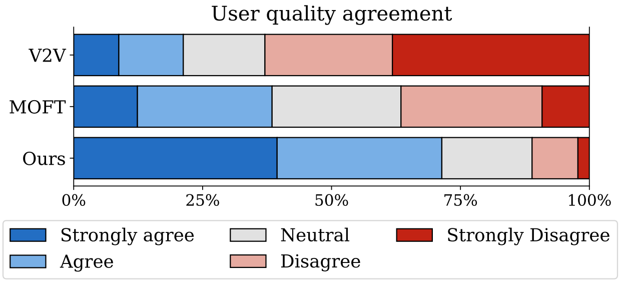
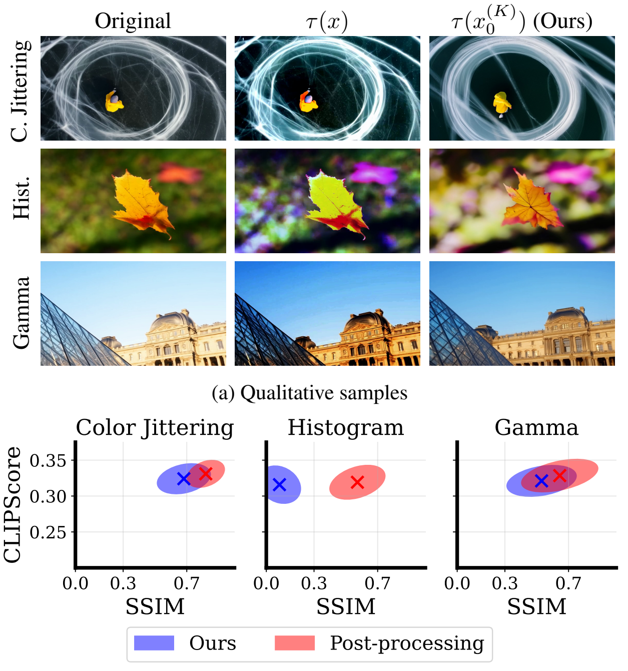


Figure 7. **User study.** We evaluate users' agreement with a statement describing match cuts, to assess how much our generated videos align with the requirements in terms of visual consistency and prompt adherence. We significantly outperform all baselines.

both structure and motion across scenes. In the first row, the burning flame transitions into the sunrise reflection, creating a transition that aligns well with the match-cut effect.

Metrics evaluation. Note that match-cut synthesis is a novel task, making quantitative evaluation inherently challenging. Nonetheless, we present quantitative results in Table 1. We start from a lower-bound baseline (T2V), defined as prompting CogVideoX-5B independently for (ρ', ρ''). The lower bound achieves a moderate CLIPScore due to its performance as a T2V method, but fails to capture continuity across scenes, as reflected by its low Motion Consistency (**0.40**) and high LPIPS (**0.74**). In contrast, V2V achieves the lowest LPIPS (**0.31**) and the highest Motion Consistency (**0.71**), indicating strong structural alignment across frames as expected. However, its CLIPScore is significantly lower than that of all other methods, suggesting the difficulty to adhere to highly distinct prompts, as seen in Fig. 6. Conversely, motion transfer methods introduce too much freedom in the scene structure, as confirmed by the considerably higher LPIPS (**0.73** for SMM, **0.53** for MOFT). Finally, MatchDiffusion enjoys a well-balanced performance. With a CLIPScore of **0.33**, it matches the prompt adherence of SMM, MOFT and the vanilla model (T2V), while still achieving a Motion Consistency (**0.69**) and LPIPS (**0.32**) score that matches the V2V baseline.

User study. Match-cuts target human audiences, and thus we conduct an evaluation against baselines based on user quality assessment. In this evaluation, we aim to quantify the smoothness of our transitions, while respecting different prompts. To do so, we show users both prompts, (ρ', ρ''), along with the match-cuts generated by MatchDiffusion and baselines. We then ask users to evaluate their agreement in a Likert-5 [21] scale with: *"This video accurately reflects the scenes described by the text and smoothly transitions between them, maintaining consistent colors, structure, movement, and appearance from one scene to the next one"*. This question assesses if the videos align with the expected consistency, while preserving different semantics. We query 35 users (average age 30.11±7.29 years old). We



(a) Qualitative samples

(b) Impact on CLIPScore

Figure 8. **Intervention effects.** In (a), we verify that our user intervention strategy allows to depart from the original image following τ with no detrimental impact to realism. We quantify this effect in (b): although our SSIM with reference frames is lower, we maintain very similar CLIPScore. We plot results as mean and std.

test against MOFT only for motion transfer, to maximize the questions per method presented to users. Results are reported in Fig. 7, showing that users *significantly* prefer MatchDiffusion over the baselines. In particular, we highlight that **39.44%** of them *strongly agree* with our statement, against 12.36% for the best baseline (MOFT). This evidence suggests superior quality of our match-cuts.

4.4. Evaluating user interventions

We evaluate our optional user intervention strategy (Section 3.3) to test whether MatchDiffusion can relax strict color/structure adherence while still generating match-cuts. We apply three τ functions to $x_0^{(K)}$: (1) color jittering, (2) histogram matching with random COCO [24] images, and (3) gamma correction. Ideally, τ should integrate with Disjoint Diffusion, preserving realism and scene structure. Fig. 8a shows results: applying τ post-generation exaggerates colors, reducing realism, while applying it to $x_K^{(0)}$ ("Ours") preserves realism despite minor structure shifts (*e.g.*, leaf shape). This allows transformations that maintain scene composition, making them suitable for match-cuts.

To assess the quality of these interventions, we randomize the parameters of τ five times and apply it to 36 synthesized video pairs, x' and x'', generating a total of 180 for each of the two strategies. We then compute SSIM [48] between the videos and their post-processed counterparts to quantify visual modifications, along with CLIPScore of $\tau(x)$ with the corresponding ρ. Fig. 8b shows that our method maintains the same CLIPScore while achieving a greater reduction in SSIM compared to the post-processing alternative. This suggests that our approach alters the video's appearance more significantly while still adhering to the prompt. In Fig. 8b histogram matching shows lower SSIM due to structural changes (*e.g.*, the leaf in Fig. 8a). Our experiments show that our pipeline produces diverse videos, smoothly integrating τ and enabling varied match-cuts.

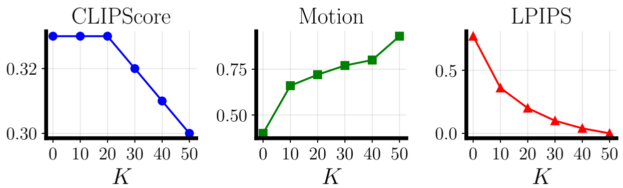


Figure 9. **Effects of K.** Increasing K to the maximum produces a hybrid video between prompts, maximizing motion fidelity and bringing LPIPS to zero. CLIPScore is slightly impacted since hybrid videos present traits of both prompts.

4.5. Impact of K

We investigate the impact of the number of Joint Diffusion steps (K) on MatchDiffusion. Fig. 9 shows the impact of K on metrics. While most results presented in Figure 5 have K between 10 and 15, we notice that although CLIPScore decreases, Motion Fidelity and LPIPS monotonically improve. This fact deserves *ad hoc* considerations. The case of $K = 0$ is equivalent to the lower bound (*i.e.* no shared structure), while $K = 50$ means that x' and x'' share *all the diffusion process* (similar to Factorized Diffusion [12]), hence $x' = x''$. In this case, MatchDiffusion produces a hybrid video (shown in supplementary). This property is not useful for match-cuts but might enable other applications. For the purpose of match-cut generation, the user's needs play a central role; K serves as a tunable parameter to adjust the results according to artistic preferences.

5. Conclusions

We presented MatchDiffusion, the first method for the synthesis of match-cuts. We formalized the match-cut generation problem as a synthesis of two videos, and proposed a method for exploiting emerging characteristics of diffusion models. MatchDiffusion has limitations, opening future research directions. Effective prompting requires creativity, and automated prompt engineering could make the method more accessible to a broader audience. Finally, refining conditioning mechanisms to give users control over specific aspects of the match-cut could simplify interaction and reduce reliance on precise prompts.

Aknowledgements The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940. We thank the reviewers for their constructive feedback, in particular for suggesting alternative transition strategies between the two videos.

References

[1] Adobe Creative Cloud. What is a match cut?, n.d. Accessed: 2024-11-14. 2

[2] S Alireza Golestaneh and Lina J Karam. Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In *CVPR*, 2017. 3

[3] Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*, 2024. 1

[4] Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szpektor, Aditya Grover, and Kai-Wei Chang. Talc: Time-aligned captions for multi-scene text-to-video generation. *arXiv preprint arXiv:2405.04682*, 2024. 2

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 2

[6] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *SIGGRAPH*, 2024. 3

[7] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. *arXiv preprint arXiv:2501.08331*, 2025. 2

[8] Angela Castillo, Jonas Kohler, Juan C Pérez, Juan Pablo Pérez, Albert Pumarola, Bernard Ghanem, Pablo Arbeláez, and Ali Thabet. Adaptive guidance: Training-free acceleration of conditional diffusion models. *arXiv preprint arXiv:2312.12487*, 2023. 2, 3

[9] Boris Chen, Amir Ziai, Rebecca S Tucker, and Yuchen Xie. Match cutting: Finding cuts with smooth visual transitions. In *WACV*, 2023. 2

[10] Dennis Fedorishin, Lie Lu, Srirangaraj Setlur, and Venu Govindaraju. Audio match cutting: Finding and creating matching audio transitions in movies and videos. In *ICASSP*, 2024. 2

[11] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *ECCV*, 2024. 2

[12] Daniel Geng, Inbum Park, and Andrew Owens. Factorized diffusion: Perceptual illusions by noise decomposition. In *ECCV*, 2024. 2, 3, 8

[13] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *CVPR*, 2024. 2, 3

[14] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 2

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3

[17] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 5

[18] Jaihoon Kim, Juil Koo, Kyeongmin Yeo, and Minhyuk Sung. Synctweedies: A general generative framework based on synchronized diffusions. *arXiv preprint arXiv:2403.14370*, 2024. 2

[19] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM Trans. Graph.*, 36(4):130–1, 2017. 2

[20] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36:50648–50660, 2023. 2

[21] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 7

[22] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023. 2

[23] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, 2024. 2, 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 8

[25] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 2

[26] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*, pages 468–485. Springer, 2024. 2

[27] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2, 4, 5, 1

[28] Walter Murch. *In the Blink of an Eye*. Silman-James Press Los Angeles, 2001. 2

[29] No Film School. How to use match cuts to tell stories, n.d. Accessed: 2024-11-14. 2

[30] Alejandro Pardo, Fabian Caba, Juan León Alcázar, Ali K Thabet, and Bernard Ghanem. Learning to cut by watching movies. In *ICCV*, 2021. 2

[31] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. Moviecuts: A new dataset and benchmark for cut type recognition. In *ECCV*, 2022. 2

[32] Sen Pei, Jingya Yu, Qi Chen, and Wozhou He. Automatch: A large-scale audio beat matching benchmark for boosting deep learning assistant video editing. *arXiv preprint arXiv:2303.01884*, 2023. 2

[33] Alonso Perez. My favourite match cut, n.d. Accessed: 2024-11-14. 2

[34] Mariano Prunes, Michael Raine, and Mary Litch. Film analysis guide. *New Haven, CT*, 2002. 2

[35] Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. Boosting diffusion models with moving average sampling in frequency domain. In *CVPR*, 2024. 2, 3

[36] 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. 2

[37] Vasco Ramos, Yonatan Bitton, Michal Yarom, Idan Szpektor, and Joao Magalhaes. Contrastive sequential-diffusion learning: An approach to multi-scene instructional video synthesis. *arXiv preprint arXiv:2407.11814*, 2024. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[40] Yaojie Shen, Libo Zhang, Kai Xu, and Xiaojie Jin. Autotransition: Learning to recommend video transition effects. In *ECCV*, 2022. 2

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 5

[42] StudioBinder. Match cuts: Creative transitions examples, n.d. Accessed: 2024-11-14. 2

[43] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *T-IP*, 2018. 3

[44] VEGAS Creative Software. Types of match cuts, examples, and how to use them, n.d. Accessed: 2024-11-14. 2

[45] 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. In *ICLR*, 2024. 2

[46] Fu-Yun Wang, Zhaoyang Huang, Qiang Ma, Guanglu Song, Xudong Lu, Weikang Bian, Yijin Li, Yu Liu, and Hongsheng Li. Zola: Zero-shot creative long animation generation with short video model. In *ECCV*, 2024. 2

[47] 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*, 2024. 1

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *T-IP*, 2004. 8

[49] 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. 3

[50] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *arXiv preprint arXiv:2405.14864*, 2024. 2, 5, 1

[51] Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788*, 2024. 2

[52] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, 2023. 5

[53] 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. In *ICLR*, 2025. 3, 5

[54] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *CVPR*, 2024. 2, 5, 1

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

[56] Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8671–8681, 2024. 2

[57] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. In *ICCV*, 2023. 2