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

Pathways on the Image Manifold: Image Editing via Video Generation

Noam Rotstein Gal Yona Daniel Silver
Roy Velich David Bensaïd Ron Kimmel

Technion - Israel Institute of Technology

Abstract

Recent advances in image editing, driven by image diffusion models, have shown remarkable progress. However, significant challenges remain, as these models often struggle to follow complex edit instructions accurately and frequently compromise fidelity by altering key elements of the original image. Simultaneously, video generation has made remarkable strides, with models that effectively function as consistent and continuous world simulators. In this paper, we propose merging these two fields by utilizing image-to-video models for image editing. We reformulate image editing as a temporal process, using pretrained video models to create smooth transitions from the original image to the desired edit. This approach traverses the image manifold continuously, ensuring consistent edits while preserving the original image's key aspects. Our approach achieves state-of-the-art results on text-based image editing, demonstrating significant improvements in both edit accuracy and image preservation. Visit our project page.

1. Introduction

Image editing has witnessed remarkable advancements through deep learning and text-guided diffusion networks. These developments have set a new benchmark for image manipulations, enhancing both control and quality. However, current approaches continue to face significant limitations in real-world scenarios. These methods often struggle with two key challenges: achieving precise edits that accurately reflect the intended modifications, and preserving the essential characteristics of the original image content.

State-of-the-art techniques predominantly rely on text-guided diffusion models, which iteratively denoise random latent representations to generate edited images. Such methods condition the generation process on both the source image—using techniques such as latent inversion [32] or model fine-tuning [22]—and the target edit description. However, these approaches require the model to generate a single output image that preserves source image fidelity while implementing complex edits, often compromising edit accuracy and content preservation.

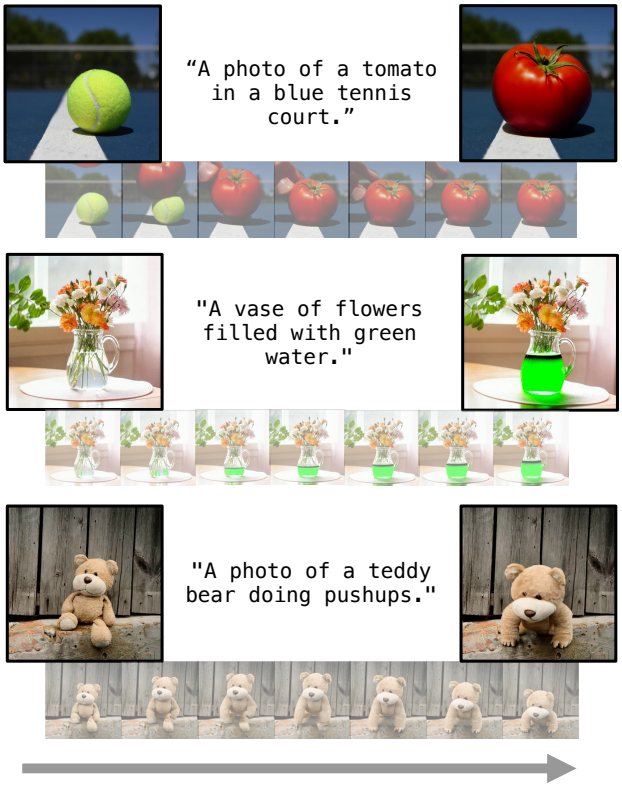


Figure 1. **Visualization of Frame2Frame's editing process.** Temporal progression of our video-based approach. Starting from the source image (leftmost), frames illustrate the natural evolution toward the target edit (rightmost). Our method produces temporally coherent intermediate states while preserving fidelity to both the source content and the editing intent.

In this paper, we propose a paradigm shift in image editing by reformulating it as a video generation task. Rather than a single-state transition, our approach harnesses temporal coherence: the source image serves as the initial frame of a video that progressively and naturally transforms toward the target edit. This temporal evolution allows the editing process to unfold through physically plausible intermediate states, providing a continuous path between source

and target images, as illustrated in Figure 1 and Figure S9 in the appendix. This temporal approach leverages the sophisticated world understanding embedded in recent video generation models, which have achieved breakthrough results in temporal coherence and visual quality through training on large-scale internet data [2, 48]. From a geometrical perspective, conventional editing approaches project initial noise onto the natural image manifold, targeting a single point where the image aligns with both the source and the edit request. In contrast, our approach generates a continuous path along the manifold between the original and edited image, producing a smooth realistic transition across different image states, as thoroughly discussed in Section 4.

We implement the proposed approach through a structured pipeline called **Frame2Frame** (F2F). First, we transform the edit instruction into a *Temporal Editing Caption* — a scenario describing how the edit should naturally evolve over time—using a pretrained Vision-Language Model (VLM). Next, a state-of-the-art image-to-video model generates a temporally coherent sequence guided by the temporal caption. Finally, we identify the frame that best realizes the desired edit with the assistance of a VLM. Extensive experiments demonstrate improvements over existing image-to-image approaches. We evaluate on TEdBench [22] and PosEdit, a newly curated dataset derived from UTD-MHAD [6], which focuses on human pose transformations. PosEdit pairs source images with ground-truth targets of the same subject in different poses, enabling rigorous evaluation of both edit accuracy and source fidelity. Beyond commonly defined editing tasks, our framework shows promising results in more classical computer vision problems such as de-blurring, de-noising, and relighting by recasting them as temporal progressions, suggesting broader applications for video-based image transformations.

Our main contributions include:

1. Reformulating image editing as a generative video task–leveraging temporal coherence to create edit paths on the natural image manifold, enabling high-fidelity manipulations while preserving source characteristics.
2. Frame2Frame: an end-to-end framework that realizes the reformulation through three key components: (1) temporal editing captions, (2) generated video-based editing, and (3) automated frame selection.
3. Comprehensive evaluation showing state-of-the-art performance on TEdBench and PosEdit, a new dataset for evaluating human pose edits.

2. Related Efforts

2.1. Image Editing

Text-based image editing has advanced significantly with the success of generative diffusion models [19]. These models, in their text-to-image version, generate images through a denoising diffusion process conditioned on input text, relying on ground truth text-image pairs for training [37, 39]. In contrast, image editing often lacks predefined ground truth data for source and target images, posing unique challenges. This limitation has led researchers to explore diverse editing methodologies [20]. For example, SDEdit [30] injects noise into an image and then denoises it based on an editing target prompt. Imagic [22] fine-tunes a text-to-image model on a single image, subsequently interpolating between input and target text embeddings to produce edits. Other methods first invert the input image into a diffusion model's latent space [21, 32] and then generate the edited image from that latent representation using various techniques for structure preservation and manipulation [18, 34]. InstructPix2Pix [5] synthesizes an editing dataset using the approach in [18], filters it, and employs this dataset to train a diffusion model in a supervised fashion. Paint-by-Inpaint [47] further explores this supervised approach, generating a real-image dataset for object insertion.

2.2. Generative Video Models

Recent years have seen remarkable advancements in video generation, evolving from domain-specific systems [15] and brief clips to models capable of generating diverse, high-fidelity content. This progress has stemmed from paradigm shifts and the large-scale expansion of datasets [7] and architectures. Approaches have evolved from recurrent networks [16, 42] and generative adversarial networks (GANs) [9, 43, 44] to latent diffusion models (LDMs) [2, 3, 11], which leverage large U-Net or transformer-based architectures alongside vast internet-sourced datasets. Notable recent efforts include Stable Video Diffusion [2], which trains an LDM on large curated datasets, and OpenAI's Sora [28], which achieves impressive results by integrating large architectures with extensive public and private datasets. These models, often termed "world simulators" due to their emergent understanding of physical dynamics and temporal coherence. Our work builds upon CogVideoX [48], a transformer-based latent diffusion model that employs a 3D Variational Autoencoder to compress videos across both spatial and temporal dimensions, enhancing coherence. To improve text-video alignment, CogVideoX also integrates an expert transformer with adaptive LayerNorm, enabling deep fusion of visual and textual modalities.

Close to our work, several recent efforts have utilized the world simulation capabilities of video diffusion models for various computer vision tasks. Make-A-Video3D [41] temporally extends static NeRFs using Score Distillation Sampling from video models. ViVid-1-to-3 [25] generates images along a camera trajectory around objects to enable novel view synthesis. PhysDreamer [51] models rigid object properties through 3D Gaussians and material fields, trained via distillation from pretrained video generators.

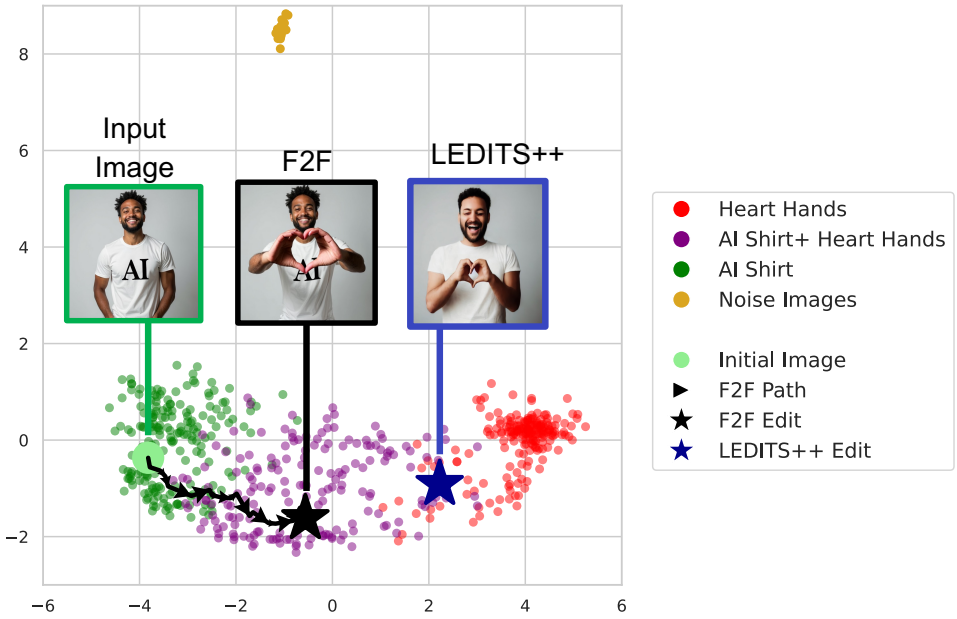


Figure 2. **Editing Manifold Pathway.** Given an input image and target caption "A happy person making a heart shape with their hands", our method generates a continuous path on the natural image manifold. Each generated frame (indicated by black arrows) represents a plausible intermediate state between the source and target, maintaining temporal consistency throughout the transformation. As a result, in contrast to the competing approach, F2F achieves the desired edit while preserving the "AI" text on the person's shirt.

2.3. Image Editing and Video

The intersection of image editing and video has received limited attention. Existing methods focus on sampling pairs of random frames from videos to build image pair datasets that capture the same subject under varying conditions. For example, AnyDoor [8] uses the paired frames as an augmentation method, segmenting foreground objects in each frame and assigning one masked object as the target edited appearance of the subject. MagicFixup [1] employs these frames to build a dataset focused on refining user-made subject coarse 2D edits. Recently, drag-based editing approaches [29, 40], used frames extracted from video and computed optical flow to collect a dataset aimed at editing by spatially dragging points within the image. In contrast to these approaches, which rely on frame sampling for image data collection, we are the first to directly perform image editing using generative video models.

3. Frame2Frame

We present Frame2Frame, a framework that reformulates image editing as a temporal transformation process. Our approach leverages video generation models to create natural transitions between source and target images, achieving consistent and realistic edits. The proposed method has three main steps, as illustrated in Figure 3.

3.1. Temporal Editing Captions

Text-based image editing methods typically operate on two inputs: a source image I_s and a target caption c, where c specifies the desired modifications to I_s. Our approach differs fundamentally by modeling editing as a temporal process. This requires a novel type of prompt—the *Temporal Editing Caption*, denoted by $\tilde{c}$—that describes the sequential transformation from source to target image. We construct $\tilde{c}$ by combining information from I_s and c to create a description of how the desired edit unfolds over time.

To automate this process, we leverage recent advances in Vision-Language Models (VLMs) [10, 14, 38, 45, 46]. The VLM, specifically ChatGPT-4o [33], is instructed to produce a concise video scenario that highlights how elements within the image change or move over time. The generated caption captures the essential transformations while maintaining a static camera perspective unless movement is necessary. To improve generation quality, we utilize in-context learning (ICL) [12], providing the VLM with nine exemplar prompt-caption pairs. The complete prompt template, ICL examples, and an ablation study comparing this approach with directly using the target captions are included in Section B of the Appendix.

3.2. Video Generation

We employ CogVideoX (I2V-5B) [48], a pretrained generative video latent diffusion model utilizing a transformer-based architecture. Specifically, we use its image-to-video variant, which has been fine-tuned to generate videos starting from an input image I_s. During generation, I_s is encoded and concatenated with noise in latent space, where

the model applies a denoising process guided by the temporal caption $\tilde{c}$. As elaborated in Section 4, this conditioning allows generated videos to start from I_s and evolve naturally along the image manifold, maintaining temporal coherence and consistency. Additionally, the model's transformer architecture enables effective fusion of visual and textual information, allowing precise control over the editing process through our temporal captions. Formally, given the video generator G, we define the generation process as:

$$\mathrm{G}(I_s, \tilde{c}) = V = \{f_1, \ldots, f_T\}$$

where V denotes the generated video with T frames, and f_t represents the frame at timestep t.

3.3. Frame Selection

We observed that the optimal number of frames required for an edit can vary—small changes may be completed in fewer frames, while more extensive transformations often necessitate additional ones. Additionally, later frames tend to deviate further from the source image. Thus, even though V serves as an editing path originating from I_s, there is no guarantee that f_T is the optimal edited image in V. For instance, in Figure 3, the edited image of the cat might be expected to capture the midpoint of its jump, but the last frame shows the jump already completed. Therefore, we aim to identify the optimal edited frame, denoted f_{t^*}, which corresponds to the earliest timestep t that achieves the desired edit. The transition from the initial frame f_1 (or I_s) to the final edited frame f_{t^*} motivates our method's name, *Frame2Frame*.

To automate the selection of t^* and avoid manual frame-by-frame review, we employ an automated approach. After generating the sequence V, we sample every fourth frame, imprinting each with a unique identifier and assembling them into an image collage alongside I_s. Inspired by [23], which introduces a novel approach to video comprehension by transforming videos into image grids, we use a VLM, specifically GPT-4o, to assist in selecting t^* by providing it with the collage and the editing prompt c. The VLM is tasked with identifying the frame that best fulfills the editing intent, evaluating each frame's alignment with c and fidelity to I_s. The model is instructed to select the optimal frame with the lowest index that completes the edit. An ablation study, detailed in Section C of the appendix, evaluates the effectiveness of our automated approach against the baseline of selecting the final frame of the video.

Note that while we refer to an optimal t^*, the definition of the required edit can be subjective and dependent on user preferences. For example, in Figure 3, different users may select frames of the cat jumping based on variations in its altitude. Thus, frame selection could serve as a flexible and customizable advantage in certain scenarios.

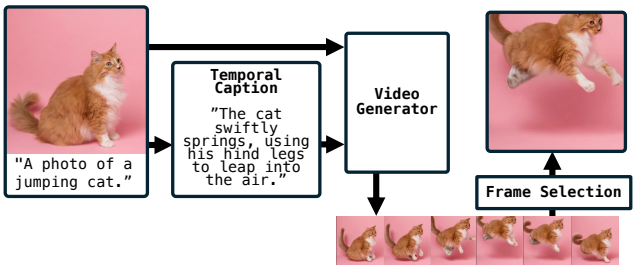


Figure 3. **Frame2Frame Overview**. Given a source image and editing prompt, our pipeline proceeds in three steps. First, a Vision-Language Model generates a temporal caption describing the transformation. Next, this caption guides a video generator to create a natural progression of the edit. Finally, our frame selection strategy identifies the optimal frame that best realizes the desired edit, producing the final image of the cat mid-leap.

4. Editing Manifold Pathway

To illustrate the advantages of our method over conventional image-to-image approaches, we aim to visualize the editing process within the natural image manifold, where realistic images reside. Given the high dimensionality of this space, we project it into a lower-dimensional representation suitable for interpretation. To achieve this, we first generate three sets of images, each containing 200 samples, using the text-to-image generator FLUX.1-dev [26]. The sets are based on the following prompts:

1. **AI**: *"Full-body portrait of a happy person wearing a shirt with the word 'AI' on it"*.
2. **AI + Heart Hands**: *"Full-body portrait of a happy person wearing a shirt with the word 'AI' on it and making a heart shape with their hands"*.
3. **Heart Hands**: *"Full-body portrait of a happy person making a heart shape with their hands"*.

We manually filter the outcomes to ensure alignment with descriptions, with examples provided in the supplementary material in Section D. Then, we use a CLIP ViT-B/32 model [36] to extract a 512-dimensional feature vector for each image. Using these features, along with 25 feature vectors extracted from random noise images, we perform Principal Component Analysis (PCA) to reduce the dimensionality into a two-dimensional subspace. The resulting subspace, now suitable for visualization, is depicted in Figure 2.

As illustrated in the figure, the natural images form a smooth manifold, distinctly separated from the distribution of noise images. Within this manifold, there is a clear semantic progression: images of people with 'AI' shirts (green cluster) are close to images of people with 'AI' shirts making a heart shape (purple cluster), which are adjacent to images of people only making a heart shape (red cluster). Thus, transitioning smoothly along the manifold allows a person with an 'AI' shirt to perform a heart shape with their hands while preserving the shirt's text.

Model	Source		Target
	LPIPS$_\downarrow$	CLIP-I$_\uparrow$	CLIP$_\uparrow$
SDEdit	0.30	0.85	0.60
Pix2Pix-ZERO	0.29	0.84	0.62
Imagic	0.52	0.86	**0.63**
LEDITS++	0.23	0.87	**0.63**
FlowEdit	**0.22**	**0.89**	0.61
F2F	**0.22**	**0.89**	**0.63**

Table 1. **TEdBench Results.** Quantitative comparison on TEdBench benchmark. Source metrics (LPIPS and CLIP-I) measure content preservation, while Target metric (CLIP) measures edit accuracy. Our Frame2Frame (F2F) method achieves better or comparable performance across all metrics.

Model	Source		Target		
	LPIPS$_\downarrow$	CLIP-I$_\uparrow$	LPIPS$_\downarrow$	CLIP-I$_\uparrow$	CLIP$_\uparrow$
GT (Reference)	0.08	0.91	0	1	0.61
SDEdit	0.39	0.61	0.39	0.64	0.57
Pix2Pix-ZERO	0.39	0.57	0.40	0.60	0.56
LEDITS++	0.26	0.65	0.28	0.69	**0.64**
F2F	**0.14**	**0.82**	**0.15**	**0.84**	**0.64**

Table 2. **PosEdit Results.** Quantitative evaluation on PosEdit. Source metrics assess similarity to the original image, while Target metrics include LPIPS and CLIP-I comparisons to the ground-truth target image, along with the CLIP score for edit accuracy. "GT" provides ground-truth target image metrics for context.

Consider an original image from the AI group that we wish to edit using the target prompt: "A happy person making a heart shape with their hands". Current editing methods generate a single image, which may cause an abrupt transition to the red cluster, effectively removing the 'AI' on the shirt. This behavior is illustrated in the figure, where the edit by LEDITS++ is positioned near the red cluster, and the 'AI' text on the shirt disappears. In contrast, our method leverages video generation to perform the edit smoothly, moving along the manifold with incremental changes until reaching the required edit in the purple cluster, as indicated by black arrows in Figure 2. The temporal editing caption guiding this process is: "A happy person very slowly raising their hands to form a heart shape". This gradual progression lets us reach the purple cluster, resulting in an edited image where the person maintains the 'AI' on their shirt while making a heart shape—a faithful preservation of the original image's key attributes. This experiment illustrates that the proposed paradigm enables smooth traversal across the image manifold, enabling consistent edits while preserving the essential characteristics of the original image.

5. Experiments

We evaluate Frame2Frame against state-of-the-art image editing methods, including LEdits++ [4], SDEdit [31], Pix2Pix-Zero [35] Imagic [22], and FLUX-based FlowEdit [24]. Our evaluation spans two benchmarks: the established TEdBench [22] for general image editing, and our newly introduced PosEdit, specifically designed for human pose editing. Finally, we conduct a human evaluation to assess our method based on real user preferences.

5.1. Evaluation Protocol

We conduct our experiments following a consistent evaluation protocol across all methods and benchmarks. Following common practice [4, 22], for each method and source image, we manually select the best result from fifteen random seeds based on visual quality and edit accuracy, ensuring the same seed set is used across all methods. For all methods, we use the default hyperparameters and settings as provided in their official implementations or official Hugging Face repositories [1].

Image Preprocessing and Generation Details. We use CogVideoX [48] as our video generation backbone, which operates at a fixed resolution of 720×480 pixels. Both TEdBench and PosEdit benchmarks consist of images with 1:1 aspect ratio. To accommodate CogVideoX's resolution requirements while preserving image content, we first resize all source images to 480×480 pixels, then horizontally pad them with black pixels to reach the required 720×480 resolution (adding 120 pixels on each side). After generating the video sequence, we crop the central 480×480 region of the selected frame to remove the padding and resize it to the final evaluation resolution of 512×512 pixels, matching the standard resolution used in prior work. For video generation, we adopt the default hyperparameters proposed by CogVideoX: a guidance scale of 6, 49 generated frames per sequence, and 50 denoising inference steps. This configuration generates videos approximately 6 seconds in duration at a frame rate of 8 frames per second.

Method-Specific Requirements. Pix2Pix-Zero [35] requires additional source image descriptions along with the standard inputs. We generate these automatically using BLIP-2 [27], a state-of-the-art image captioning model, ensuring consistent source descriptions across all experiments. For our method, we transform the original editing prompts from both benchmarks into temporal editing captions, as described in Section 3.1.

5.2. TEdBench Evaluation Results

We quantitatively evaluate our method and baselines on the TEdBench benchmark using three complementary met-

[1] LEdits++, Pix2Pix-Zero, SDEdit

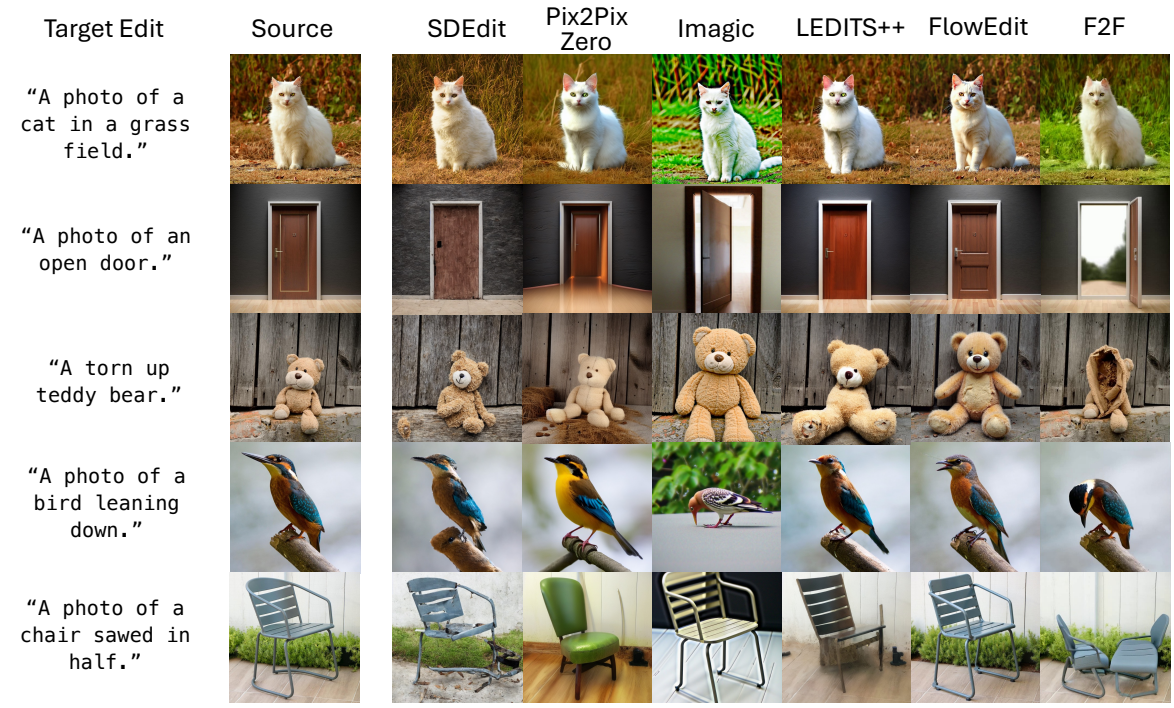


Figure 4. **Qualitative Results on TEdBench.** Comparison with other methods across various editing tasks. Our approach consistently produces edits that better align with the target prompt while preserving the source image's content and structure. For instance, in the teddy bear example, our method uniquely achieves complex structural modifications while maintaining high visual quality.

rics. For each method, we compute the metrics between the source image and its corresponding best edited version. LPIPS [50] measures perceptual similarity to assess how much the edit preserves the source image's content, with lower values indicating better preservation. CLIP-I evaluates the similarity between source and edited images in CLIP's [36] feature space, where higher values indicate better preservation of semantic content. Finally, CLIP score measures alignment between the edited image and the target prompt, with higher values indicating better adherence to the editing instruction. As shown in Table 1, our method achieves strong performance across all metrics, demonstrating effective balance between preserving source content and achieving the desired edit.

The qualitative advantages of our method are visually evident in the comparisons presented in Figure 4. The figure demonstrates our method's superior performance across diverse editing scenarios, producing results that both faithfully execute the intended edits while maintaining strong alignment with the source image. Additional examples showcasing these capabilities are provided in the appendix in Section A.

5.3. PosEdit Benchmark

We introduce PosEdit, a benchmark for human pose editing derived from the UTD-MHAD dataset [6]. UTD-MHAD includes RGB videos of 8 subjects (4 females and 4 males) performing predefined actions in a controlled indoor environment. We carefully curated 58 editing tasks encompassing 8 distinct action categories, ranging from simple poses like a raised hand to complex athletic poses such as basketball shooting and lunging.

The proposed benchmark specifically focuses on human pose manipulation. Unlike TEdBench, each editing task in PosEdit includes a ground truth frame extracted from the same subject in the target pose. The availability of reference images allows for a more comprehensive evaluation of both editing accuracy and identity preservation. In the dataset, each editing task consists of two images: a source image showing the subject in a neutral standing pose with arms relaxed at their sides, and a ground-truth edited target image capturing the subject performing in a specific pose. The benchmark also provides a prompt for each task that describes the target pose the subject should achieve (e.g., "A person in a basketball shooting posture.").

Benchmark Evaluation. The evaluation on PosEdit follows the same protocol established for TEdBench: generating multiple variants using nine different seeds and manually selecting the best result for each method. However, PosEdit's ground-truth target images enable additional evaluation metrics. Beyond measuring preservation of source content (via LPIPS and CLIP-I between source and edited images) and edit accuracy (via CLIP score with the prompt),

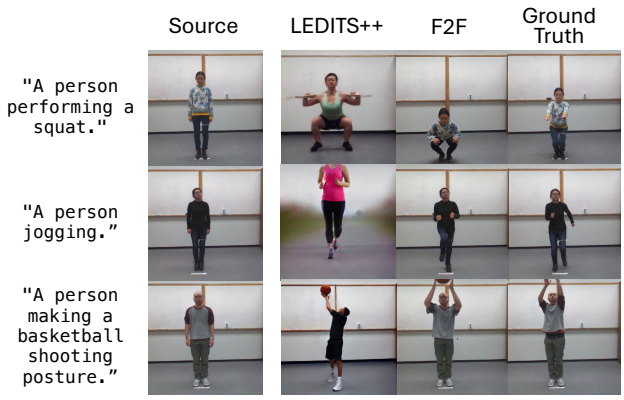


Figure 5. **Qualitative Results on PosEdit.** Comparison between our Frame2Frame method and LEDITS++ on human motion editing tasks. For each example, we show the source image, edited results from both methods, and the ground-truth target image. Our method better preserves subject identity while achieving more natural pose transitions. The evaluation metrics for each image are provided in Section F of the appendix.

we compute LPIPS and CLIP-I metrics between the edited image and its corresponding ground-truth target. This enhanced evaluation directly assesses both the accuracy of the transformation and preservation of identity features. Additionally, we include the evaluation metrics of the ground-truth images as a reference to the rest of the results.

Table 2 demonstrates that our method consistently outperforms competitive methods across all metrics, with particularly strong performance in similarity measures with the target ground-truth image. This advantage highlights our method's superior ability to preserve key features while achieving the desired edit, as illustrated by the qualitative comparisons in Figure 5. The figure shows that our method generates more natural pose transitions while maintaining crucial identity attributes such as facial features, body proportions, and clothing details.

Two important implementation notes: First, for Pix2Pix-Zero, which requires source image descriptions, we use the same static prompt across all tasks: "*A person standing naturally with his arms relaxed at his sides*." Second, we exclude Imagic [22] from this comparison as their official implementation is not publicly available and their results are only reported on TEdBench.

5.4. Human Evaluation Survey

To further evaluate our method, we conducted a human survey. As indicated in Table 1, LEDITS++ emerges as the most competitive method compared to ours, making it a natural choice for comparison. Each participant was presented with 20 randomly sampled TEdBench samples, including the source image, the target edit prompt, and the corresponding edited outputs from both methods. Inspired by the methodology proposed by [49], participants evaluated each comparison based on two criteria: (1) edit accuracy relative to the prompt, and (2) edit quality—defined as the preservation of visual fidelity to the source image, seamless integration of edited elements, and the overall natural appearance of modifications. We collected responses from 59 randomly selected online evaluators. To ensure an unbiased assessment, evaluators were not informed of the study's objectives. The results were quantified using two metrics: (i) overall global preference, expressed as a percentage, and (ii) aggregated per-image preference, where a tie resulted in each method receiving 0.5 points.

Method	**Edit Accuracy**		**Edit Quality**	
	Overall	Per-Image	Overall	Per-Image
F2F	**54.1%**	**53.0%**	**65.6%**	**67.0%**
LEDITS++	45.9%	47.0%	34.4%	33.0%

Table 3. **Human Survey Results.** Human evaluation on TEdBench shows that F2F surpasses LEDITS++ in edit accuracy while offering a significant advantage in preserving the original image.

Our results, summarized in Table 3, demonstrate that F2F outperforms LEdits++ on both metrics. Specifically: For edit quality, F2F achieved a global preference score of 53%, slightly surpassing the 47% obtained by LEdits++. In terms of edit accuracy, F2F achieved a 65.6% global preference compared to LEdits++'s 34.4%. The per-image preference results mirrored this trend, indicating robustness across individual examples and no significant influence of outliers. These findings reinforce our claim that smooth temporal editing—using video as a medium—preserves essential scene characteristics while successfully performing the edit. Our results suggest that the gap between F2F and LEdits++ in source preservation is larger than reflected by the LPIPS scores in Table 1. The full survey details are provided in the supplementary materials in Section G.

5.5. Additional Vision Tasks

We demonstrate our framework's applicability beyond traditional editing by applying it to fundamental image manipulation tasks: denoising, deblurring, outpainting, and relighting. For these experiments, we utilize Runway Gen-3 [13] as our video generation backbone, as it showed superior performance on these specific tasks compared to other generative video models. Since these tasks serve as extensions and potential applications of our method, we focus on qualitative results, leaving quantitative evaluation for future work. Our results, shown in Figure 6, demonstrate strong performance across all tasks. We associate this success to the natural alignment between these operations and common video scenarios: deblurring maps to camera focusing, outpainting to camera motion, and relighting to time-lapse lighting changes. This alignment enables our

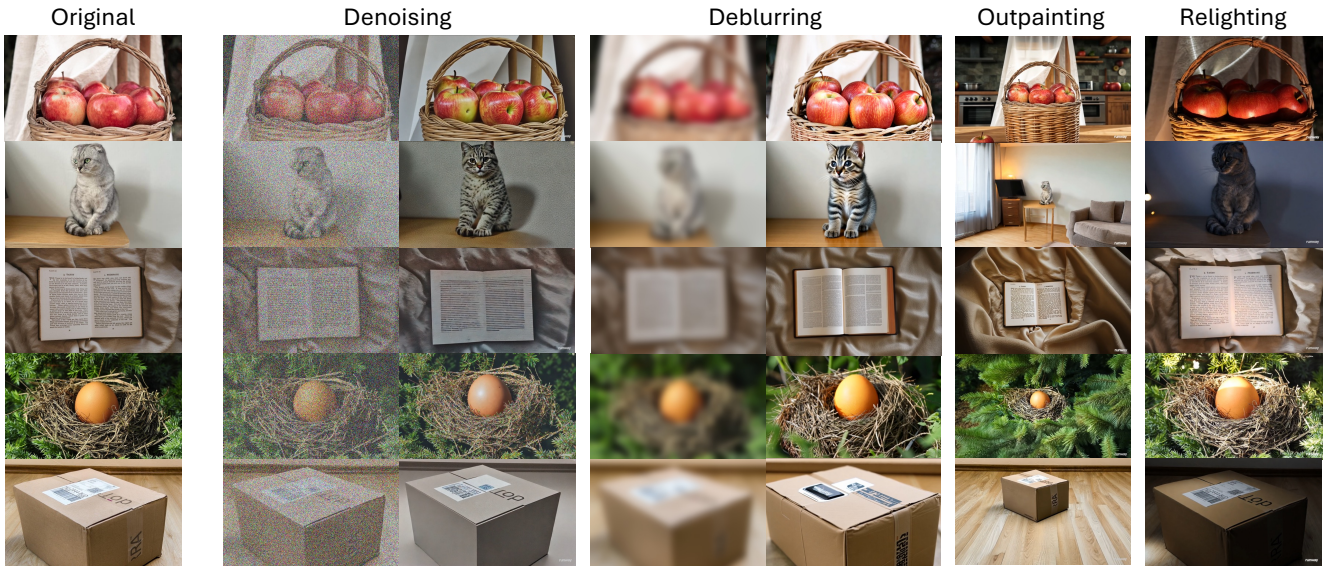


Figure 6. **Additional Vision Tasks.** Qualitative results of our image-to-video-to-image editing approach on selected traditional tasks.

method to leverage the video model's learned temporal dynamics, achieving high-quality results without task-specific training. Full prompts are provided in Section I of the supplementary material.

Extended Results. Additional examples illustrating our method's editing performance across diverse backgrounds are provided in Section H of the Appendix.

6. Limitations

While our approach addresses key shortcomings in current image editing methods, it also introduces unique challenges. For example, natural camera motion sometimes appears in video sequences, which, when replicated in generated content, can lead to unintended perspective shifts. As with all generative models, video models are trained on specific data domains, making it challenging to produce results that deviate significantly from the model's training data, which predominantly includes real-world transformations. Despite these challenges, F2F demonstrates success in several cases. For example, in the middle row of Figure 1, the vase is "magically" filled with green water, showcasing the model's ability to perform imaginative edits, with additional examples discussed in Section H of the supplementary. Additionally, our method is computationally intensive, as transforming an image into a video sequence is resource-heavy and often slower than other image editing methods. However, video generation efficiency is advancing rapidly, as demonstrated by models like Runway Turbo[2] and LTX-Video [17], which can execute the process in seconds, making this approach increasingly less resource-intensive. Additionally, optimizing the number of frames per edit—currently fixed at 49 for CogVideoX—could enable faster editing possibilities.

[2] Runway Turbo API

7. Conclusions

We introduced Frame2Frame, a novel approach that reformulates image editing through video generation. By leveraging video models' inherent understanding of temporal transformations, our method achieves state-of-the-art editing results while maintaining high fidelity to source images. We demonstrated our framework's effectiveness on standard benchmarks, introduced PosEdit for human pose editing, and showed promising results on more classical vision tasks. As video generation technology advances, we expect our approach to enable increasingly sophisticated image manipulations while maintaining natural and physically plausible results.

Future Research. Our work opens several promising directions. First, fine-tuning video generators specifically for image editing presents an exciting opportunity, including straightforward solutions like enforcing static camera scenarios or using datasets curated for editing tasks, alongside more complex approaches yet to emerge. Secondly, a key direction could involve reducing the overhead of full video generation while preserving the benefits of gradual, temporally coherent transformations, potentially enhancing both the efficiency and speed of editing. The speed can be affected both in terms of per-frame average generation speed or the number of frames generated per edit, which is currently fixed at 49 for CogVideoX.

References

[1] Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. Magic fixup: Streamlining photo editing by watching dynamic videos. arXiv preprint arXiv:2403.13044, 2024. 3

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023. 2

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22563–22575, 2023. 2

[4] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. 2023. 5

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18392–18402, 2023. 2

[6] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In 2015 IEEE International conference on image processing (ICIP), pages 168–172. IEEE, 2015. 2, 6

[7] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13320–13331, 2024. 2

[8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6593–6602, 2024. 3

[9] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. In ICCV, 2019. 2

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3

[11] Duolikun Danier, Fan Zhang, and David Bull. Ld-mvfi: Video frame interpolation with latent diffusion models. In Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 2

[12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. arXiv preprint arXiv:2301.00234, 2022. 3

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 7

[14] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13861–13871, 2024. 3

[15] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5609–5619, 2023. 2

[16] David Ha and Jürgen Schmidhuber. World models. arXiv preprint arXiv:1803.10122, 2018. 2

[17] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103, 2024. 8

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020. 2

[20] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. arXiv preprint arXiv:2402.17525, 2024. 2

[21] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12469–12478, 2024. 2

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6007–6017, 2023. 1, 2, 5, 7

[23] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. arXiv preprint arXiv:2403.18406, 2024. 4, 12

[24] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. arXiv preprint arXiv:2412.08629, 2024. 5

[25] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6775–6785, 2024. 2

[26] Black Forest Labs. Flux.1-dev. https://github.com/black-forest-labs/FLUX.1-dev, 2024. 4

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023. 5

[28] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177, 2024. 2

[29] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8217–8227, 2024. 3

[30] 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. arXiv preprint arXiv:2108.01073, 2021. 2

[31] 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 International Conference on Learning Representations, 2022. 5

[32] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6038–6047, 2023. 1, 2

[33] OpenAI. Chatgpt-4o (october 2024 version). https://chat.openai.com/chat, 2024. 3

[34] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In ACM SIGGRAPH 2023 Conference Proceedings, pages 1–11, 2023. 2

[35] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In ACM SIGGRAPH 2023 Conference Proceedings, New York, NY, USA, 2023. Association for Computing Machinery. 5

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021. 4, 6

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022. 2

[38] Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models for enriched fused image captions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 5689–5700, 2024. 3

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35:36479–36494, 2022. 2

[40] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent YF Tan, and Jiashi Feng. Instadrag: Lightning fast and accurate drag-based image editing emerging from videos. arXiv preprint arXiv:2405.13722, 2024. 3

[41] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. arXiv preprint arXiv:2301.11280, 2023. 2

[42] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In International conference on machine learning. PMLR, 2015. 2

[43] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In CVPR, 2018. 2

[44] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In NeurIPS, 2016. 2

[45] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079, 2023. 3

[46] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024. 3

[47] Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. arXiv preprint arXiv:2404.18212, 2024. 2

[48] 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. 2, 3, 5

[49] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems, 36, 2024. 7, 14

[50] 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. 6

[51] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In European Conference on Computer Vision. Springer, 2025. 2