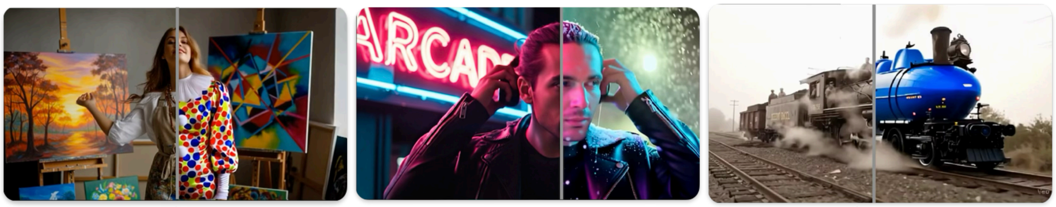
Lucy Edit: Open-Weight Text-Guided Video Editing

Decart AI Team
contact@decart.ai



"dress her like a clown" "make the scene snowy" "turn train to a rocket"

Figure 1: Editing results produced by **Lucy Edit**. For each example, the left half of the frame shows the original video and the right half shows the edited output guided by the text prompt below. Examples include local appearance change (*"dress her like a clown"*), global environmental modification (*"make the scene snowy"*), and large-scale object replacement (*"turn train to a rocket"*).

Abstract

We present **Lucy Edit**, the first open-weight foundation model for text-guided video editing. Lucy Edit enables users to modify videos directly from natural language instructions, supporting operations such as object and background replacement, clothing and accessory changes, insertion or removal of elements, and global style or scene transformations, all without requiring masks or manual annotation. This positions Lucy Edit as a unified, accessible framework for controllable video editing and a foundation for future research and applications. Achieving high-quality text-guided edits requires four properties: edits must faithfully follow the prompt, the subject's identity and fine details must be preserved, temporal consistency must hold across frames, and the results must remain realistic. Lucy Edit meets all four criteria. Our evaluations demonstrate strong identity preservation in human-centric edits, precise localization of object insertions and replacements, realistic integration of new content into motion and lighting, and strong adherence to diverse natural-language prompts. By releasing **Lucy Edit Dev** as the first open-weight foundation model in this category, alongside **Lucy Edit Pro** via API, we provide both a foundational platform for the research and creators community to build upon and a pathway to professional deployment.

1 Introduction

The ability to edit videos directly from text promises to transform how visual content is created and consumed, turning a sentence into a powerful editing tool. Text-guided video editing seeks to modify a source video according to a natural-language instruction, enabling operations such as object or background replacement, clothing and accessory changes, insertion or removal of elements, and global style or scene transformations. To be practically useful, a system have to satisfy four requirements simultaneously: *fidelity* to the prompt, *preservation* of subject identity and visual content outside the edited regions, *temporal consistency* across frames, and overall *realism* of the result. While text-to-image editing has advanced rapidly [1, 2, 3], achieving comparable quality in the video domain has remained an open challenge.

Most prior approaches fall into one of two categories. The first are *inference-time methods*, such as inversion or optimization-based pipelines. These approaches can be slow and inflexible, sensitive to hyper parameters, and prone to temporal flicker, making them impractical for longer clips or iterative use. The second are *conditional video generation models*, which rely on auxiliary inputs such as depth maps, masks, or reference images to approximate edits. While useful in constrained scenarios, these systems depend on external signals and cannot fully address the general problem of instruction-guided editing.

We introduce **Lucy Edit**, the first open-weight *foundation model* for text-guided video editing. Lucy Edit performs in-context editing by conditioning on both the input video and the edit instruction within a rectified-flow generative framework. A single, efficient modification enables a wide spectrum of edits without architectural changes, control networks, or per-clip optimization. Prompts are followed directly, edits are applied precisely, identities are preserved, and results remain realistic and temporally stable, as demonstrated in Figure 1.

Lucy Edit establishes a new baseline for high-quality text-guided video editing. Its combination of identity conservation, edit precision, realism, and prompt adherence makes it a uniquely valuable foundation model on which researchers and developers can build. As the first open-weight release in this category, and surpassing closed commercial systems in serveral aspects, Lucy Edit provides the community with a foundation model to extend and adapt. We release **Lucy Edit Dev** as an open-weight foundation for research, alongside a higher-capacity **Lucy Edit Pro** model available via API for production use, with the aim of accelerating progress toward accessible, controllable video editing for everyone.

2 Related Work

Foundation text-to-image models. Diffusion-based models revolutionized image generation by combining large language encoders with scalable denoisers. The original denoising diffusion framework [4] and its score-based counterparts [5] established the mathematical foundation for generative diffusion. Latent Diffusion Models [6, 7] (LDMs) introduced latent-space modeling for compute-efficient training. The use of pretrained language encoders such as CLIP [8] and T5 [9] proved crucial for semantic alignment between prompts and generated images, while architectures like the Diffusion Transformer (DiT) scaled generation quality through self-attention [10]. More recent refinements, such as rectified flow training [11], have further improved convergence and sampling efficiency. Together, these advances established the backbone of modern image generation, laying the groundwork for editing and video models.

Text-guided image editing. Image editing methods adapt foundation models to real inputs through prompt-space control, inversion, and instruction-following. Prompt-to-Prompt manipulates cross-attention to preserve layout while altering semantics [12]; InstructPix2Pix learns instruction-conditioned edits from paired supervision [1]; and Null-Text Inversion improves reconstruction fidelity for downstream edits [13]. Unified generation–editing systems such as FLUX.1 Kontext [14] adopt flow matching and token concatenation to support both new synthesis and local edits within one framework. Industrial efforts, including Qwen-Image-Edit [15], Seedream [16] and Google's Nano Banana, emphasize robustness and efficiency for practical workflows. As a result, image editing quality is now strong across local/global changes, identity preservation, and successive refinements, motivating analogous progress in video.

Video generation. Video generation initially extended image generation by applying spatial diffusion frame by frame, later incorporating temporal self-attention to capture motion. Pioneering models such as Imagen Video [17] and Make-A-Video [18] demonstrated the feasibility of scaling diffusion to videos by conditioning on text and leveraging frame interpolation. Subsequent work introduced dedicated spatiotemporal architectures, variational autoencoders, and hierarchical conditioning, enabling higher resolution and longer sequences [19, 20, 21, 22]. Recent large-scale models highlight the growing maturity of text-to-video synthesis, but most focus primarily on generation rather than fine-grained editing.

Text-guided video editing. Prior video editing approaches largely fall into two groups. One group relies on *inference-time methods*, including inversion, optimization, or multi-stage guidance during sampling. These approaches are flexible but computationally expensive, scale poorly to long videos, and often exhibit temporal artifacts such as flicker or drift. The second group can be described as *conditional video generation*, where auxiliary observations such as depth, masks, or reference images are supplied alongside text prompts to guide edits [23, 24, 25]. While effective for constrained scenarios, these methods are limited by their dependence on external observations and do not fully generalize to unconstrained, instruction-guided edits.

Lucy Edit differs from both categories: it performs instruction-guided editing end-to-end, requires no masks or auxiliary signals, and avoids costly inference-time optimization. By conditioning jointly on the input video and edit instruction, it achieves high-fidelity edits, preserves identity and motion, and maintains temporal stability, while being the first open-weight model in this category.

3 Method

Our goal is to edit an input video $\mathcal{V}$ according to a natural-language instruction $\mathcal{T}$, producing an output $\hat{\mathcal{V}}$ that reflects the requested edit while preserving identity, motion, and unedited regions. Lucy Edit builds on a rectified-flow text-guided video backbone and introduces a minimal but powerful modification: concatenation of the noisy latent and the encoded input-video latent along the channel dimension. This section outlines the rectified-flow backbone, our channel-concatenation strategy, and the requirements on data that make such a model possible.

3.1 Rectified Flow for Video Generation

Rectified flow [11, 26] defines a deterministic transport between Gaussian noise and the data distribution through the integration of an ordinary differential equation (ODE). A clean latent video x and Gaussian noise ε are linearly interpolated as

$$z_t = (1-t)\,x + t\,\varepsilon, \qquad t \in [0,1]. \tag{1}$$

A velocity field v_θ predicts the transport direction at time t, trained with a squared error against the ground-truth direction. During inference, the ODE is integrated from $t = 1$ (pure noise) to $t = 0$ (clean latent), producing x which is decoded into the final output video.

3.2 Editing via Channel Concatenation

To turn a generator into an editor, Lucy Edit incorporates the input video directly into the denoising process. The input clip is first encoded by the backbone VAE into latents v, while the evolving sample at time t is represented as noisy latents z_t. These are concatenated along the channel dimension before entering the denoiser, forming $\tilde{z}_t = \mathrm{concat}_{\mathrm{ch}}(z_t, v)$. Text embeddings τ are injected through cross-attention as in standard text-to-video models. Concatenation doubles the input channel dimension but leaves the dominant quadratic attention cost unchanged, yielding a negligible increase in overall runtime.

This design (depicted in Figure 2) offers both efficiency and fidelity. By enlarging only the initial projection, the computational footprint is kept minimal. At the same time, channel concatenation in latent space preserves precise spatiotemporal alignment between the conditioning video and the noisy sample. This direct alignment is especially effective for edits that demand consistency of identity and motion, such as changing clothing or inserting objects. Conceptually, inference proceeds as an alternating loop: encode the input video once, concatenate with the current noisy latent, denoise under the rectified-flow dynamics, update the latent by one schedule step, and repeat until convergence.

3.3 Requirements on Data

Training a model for text-guided video editing requires stringent properties of the data. First, the triplets of input video, edit instruction, and edited output must be perfectly aligned in time and content. Second, the instruction must correspond precisely to the visual transformation, what we term instruction–edit specificity, so that no spurious changes are learned. Third, the corpus must span both a wide range of edit types and substantial variation within each type: clothing changes must cover many garments, object insertions must include diverse categories, and scene replacements must range across styles, times of day, and lighting conditions. Finally, the training data must maintain visual

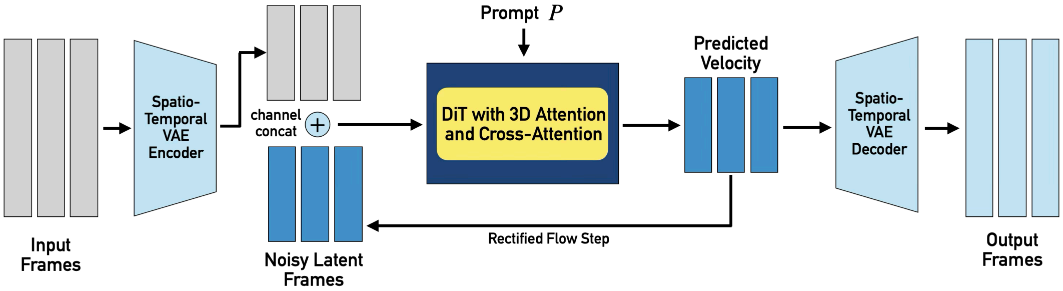


Figure 2: **Architecture Overview of Lucy Edit**. An input video is encoded into latent representations with a spatio-temporal VAE. The encoded input-video latents are concatenated along the channel dimension with noisy latent frames. This concatenated tensor, together with the text prompt P, is processed by a DiT backbone with 3D self-attention and cross-attention to predict the velocity field. At each rectified-flow step, the predicted velocity drives the noisy latents toward the clean data distribution, and the final denoised latents are decoded by the spatio-temporal VAE to produce the edited output frames.

realism and temporal coherence; synthetic corpora are often used in prior work, but over-reliance on them can lead to outputs that inherit synthetic artifacts.

Lucy Edit is trained on a heterogeneous mixture of aligned triplets, unedited captioned videos that regularize motion and appearance priors, and image–text resources extended to short video sequences via affine temporal trajectories, in the style of MovieGen [22]. Together, these sources satisfy the alignment, specificity, diversity, and quality criteria necessary for high-fidelity and temporally stable video editing.

4 Evaluations and Applications

We qualitatively evaluate **Lucy Edit** across a wide spectrum of text-guided editing tasks. Our goal in this section is not only to demonstrate that edits succeed, but also to highlight four properties that make Lucy Edit unique: (i) preservation of subject identity, (ii) precision of localized edits, (iii) realism of the resulting dynamics, and (iv) adherence to natural-language prompts. Together, these dimensions characterize what it means for a video editing model to be usable in practice.

4.1 Identity Preservation

Lucy Edit consistently preserves the identity of people and other primary subjects in a wide range of edits. Even under substantial transformations, such as clothing changes, background replacement, or stylization, facial features, body shape, and motion trajectories remain intact. This property is particularly important for human-centric applications where continuity of identity is essential.

4.2 Precision of Edits

Beyond global consistency, Lucy Edit applies edits with high spatial and semantic precision. When a prompt specifies a localized modification, such as changing hair color, changing a logo on clothing, or inserting a small object, only the relevant regions are modified while the rest of the scene remains unchanged. This allows users to trust that edits will not cause collateral alterations in unrelated areas.

4.3 Realism of Dynamics

The added or altered content is seamlessly integrated into the original video with realistic geometry, lighting, and dynamics. Clothing deforms naturally with body motion, added objects respect scene perspective, and background replacements remain temporally stable across frames. This level of realism distinguishes Lucy Edit from prior editing-by-conditioning approaches, where edits often appear pasted on or temporally unstable.

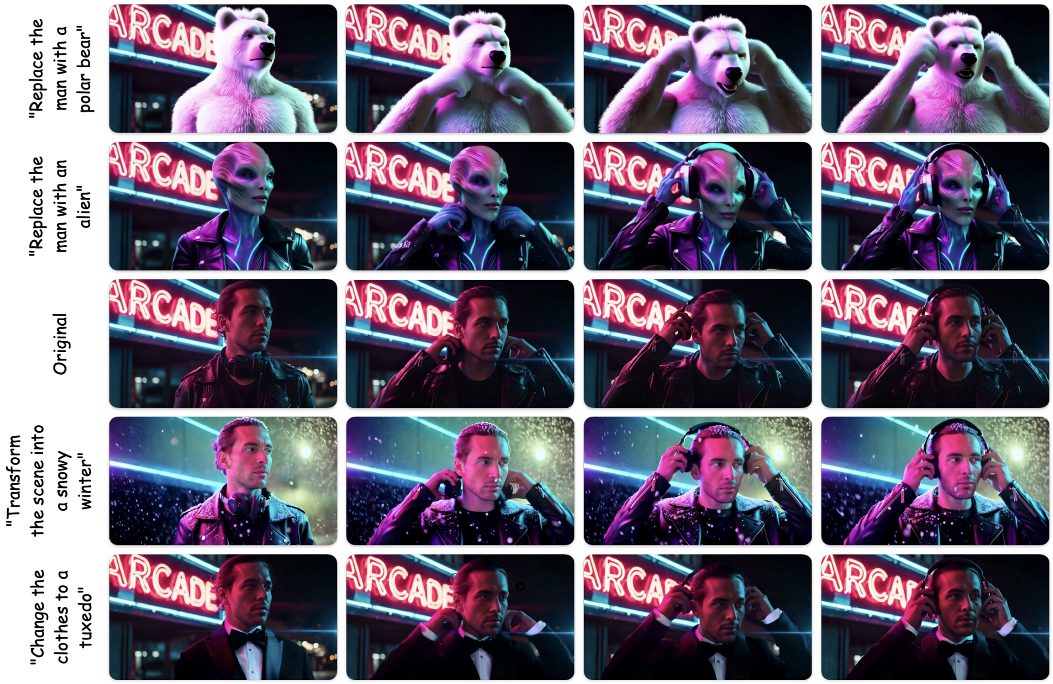

"Replace the man with a polar bear"
"Replace the man with an alien"
Original
"Transform the scene into a snowy winter"
"Change the clothes to a tuxedo"

Figure 3: **Motion Preservation.** Despite changes in outfit, character and environment, the motion remains synchronized.

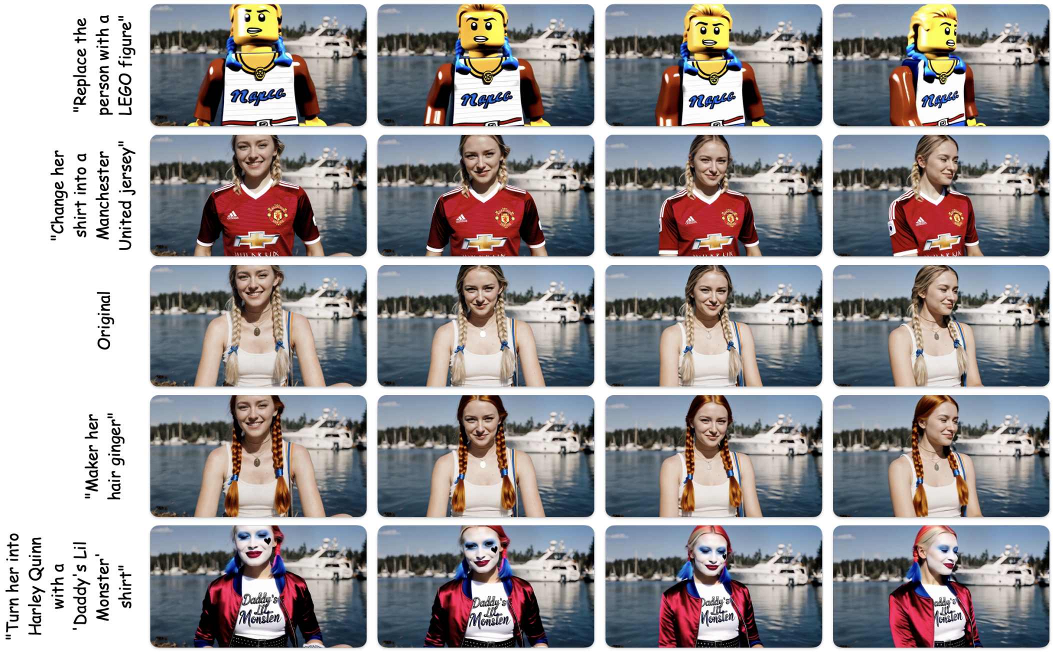

"Replace the person with a LEGO figure"
"Change her shirt into a Manchester United jersey"
Original
"Maker her hair ginger"
"Turn her into Harley Quinn with a 'Daddy's Lil Monster' shirt"

Figure 4: **Precision.** Lucy Edit modifies only the elements described by the prompt (hair color, shirt text, jersey color), leaving all other aspects of the video unchanged, such as the background, motion, etc.

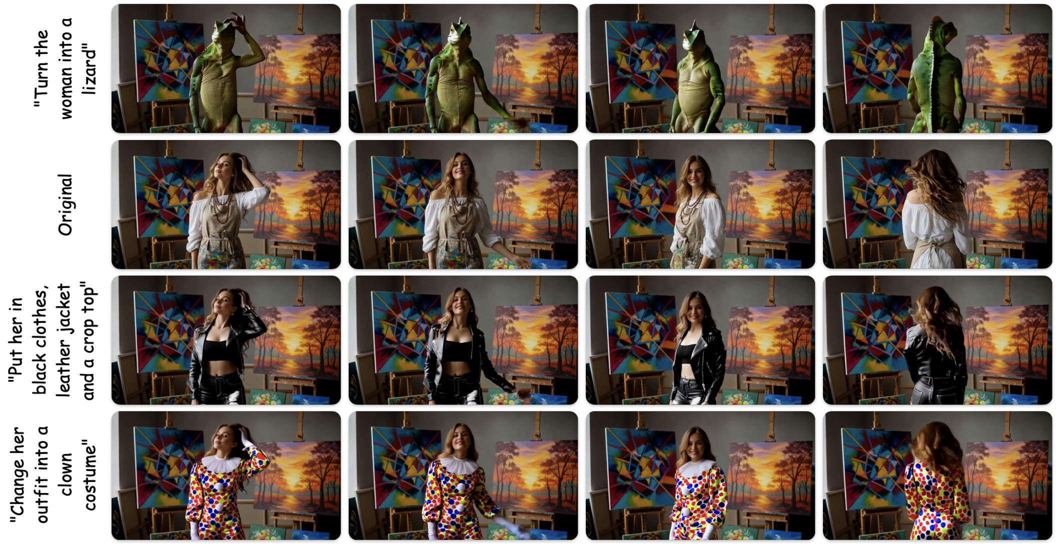

"Turn the woman into a lizard"
Original
"Put her in black clothes, leather jacket and a crop top"
"Change her outfit into a clown costume"

Figure 5: **Realism.** Clothing and accessories adapt naturally to motion, with dynamics that are physically coherent with the underlying video.

4.4 Prompt Adherence

Finally, Lucy Edit demonstrates strong adherence to user instructions. The prompts are followed directly, without the need for segmentation masks, reference images, or auxiliary conditioning inputs. This simplicity makes the system easy to use: a single sentence suffices to specify the desired transformation, and the model executes it faithfully.

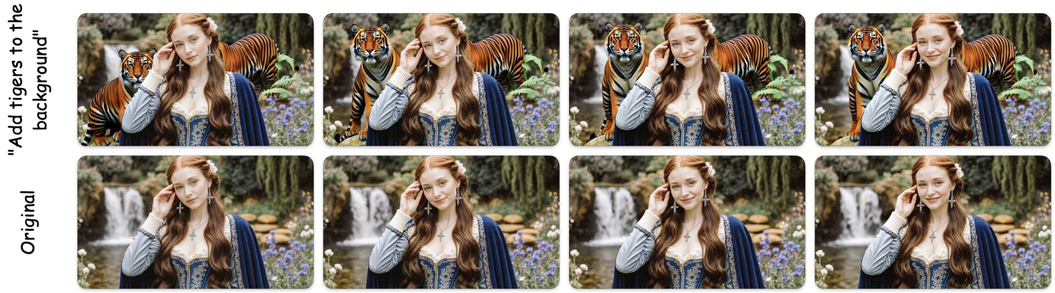


Figure 6: **Prompt adherence.** Lucy Edit inserts both dynamic and static objects (e.g., a crown, a dolphin) based solely on text, without requiring masks or external control signals.

4.5 Applications

Together, these four properties enable a wide range of downstream applications. For creative workflows, Lucy Edit allows iterative refinement of content by applying successive textual edits while retaining the subject identity. For media production, it offers a tool for altering clothing, props, or backgrounds without reshooting footage, significantly reducing the production cost. For personalization, it supports dynamic styling of avatars and characters while maintaining natural motion. The ease of use, edits specified only by text, lowers the barrier for non-experts, while the fidelity and realism make the outputs usable for professional purposes.

5 Discussion

Lucy Edit is a glimpse of the *CGI of the future*: high-fidelity video editing performed directly from text. By combining rectified-flow generation with a channel-concatenation, Lucy Edit enables edits that are faithful, temporally consistent, identity-preserving, and easy to control. Our evaluations highlight how these properties translate into practical strengths: people remain recognizable across transformations, edits are applied with surgical precision, new content integrates realistically into motion and lighting, and prompts are followed directly without auxiliary inputs. These qualities make Lucy Edit a foundation model with immediate value for creative industries, media production, and personalization.

Limitations. Despite its strengths, Lucy Edit has clear limitations. Consecutive multi-turn edits can accumulate drift, leading to artifacts or gradual degradation of temporal stability. Global or strong style transfers may alter subject identity, especially under large domain shifts such as realism to stylization. Inserted dynamic objects move plausibly with the scene but do not yet exhibit complex or intentional behaviors—for example, a newly added human will not perform a backflip.

Future Directions. Addressing these limitations suggests several promising research directions. Improving robustness to multi-turn editing would support seamless iterative workflows. Better disentanglement of global style from local identity could allow more aggressive edits without sacrificing fidelity. Learning richer motion priors could enable inserted dynamic objects to exhibit purposeful behavior. Further efficiency gains in sampling would move Lucy Edit closer to real-time interactivity.

Impact. We release both **Lucy Edit Dev**, the first open-weight *foundation model* for text-guided video editing, and **Lucy Edit Pro**, our most capable model accessible via API. Pro unlocks a new era of applications and use cases, while Dev provides the research community with an extensible platform for building new methods and applications. Together they demonstrate the core strengths of Lucy Edit, identity conservation, edit precision, realism, and prompt adherence, making it a valuable foundation for both practical deployment and scientific exploration. We hope this release accelerates the development of accessible, controllable video editing systems, ultimately bringing us closer to tools that make editing video as natural as directing a scene.

References

[1] 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] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024.

[3] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

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

[5] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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

[7] David Podell et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[11] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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

[13] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv:2211.09794*, 2022.

[14] Black Forest Labs et al. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv:2506.15742*, 2025.

[15] Qwen Team. Qwen-Image-Edit model card. Hugging Face, 2025.

[16] IndiaTimes Tech. Seedream 4.0 tutorial: Bytedance's nano banana rival, 2025.

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

[18] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

[19] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

[20] Aoxiong Yin, Kai Shen, Yichong Leng, Xu Tan, Xinyu Zhou, Juncheng Li, and Siliang Tang. The best of both worlds: Integrating language models and diffusion models for video generation, 2025.

[21] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[22] Meta AI. Movie gen: A cast of media foundation models. *arXiv:2410.13720*, 2024.

[23] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv:2503.07598*, 2025.

[24] Lin Liu et al. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv:2502.11079*, 2025.

[25] Decart AI. Miragelsd: Zero-latency, real-time, infinite controllable video generation, 2025. Technical Report.

[26] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.