Generating Past and Future in Digital Painting Processes

LVMIN ZHANG, Stanford University, USA
CHUAN YAN, Stanford University, USA
YUWEI GUO, The Chinese University of Hong Kong, Hong Kong, China
JINBO XING, The Chinese University of Hong Kong, Hong Kong, China
MANEESH AGRAWALA, Stanford University, USA

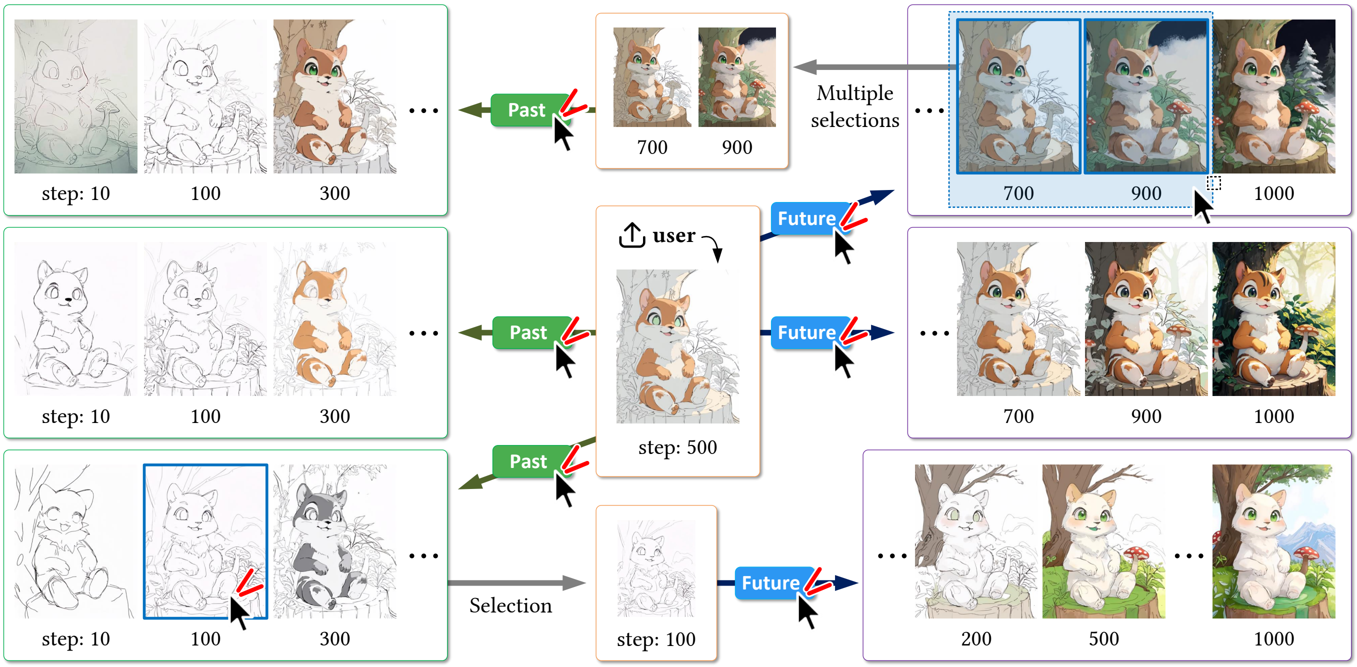


Fig. 1. Given a canvas image uploaded by user, this framework can generate states in the past and future of the drawing processes. Users can click "past" or "future" to specify the direction for the generation. They can also select one or multiple samples as inputs for generating past or future states. In this example, the central image is uploaded by the user, while all others are generated by this framework.

We present PaintsAlter, a framework to generate past and future processes for drawing process videos. Given a canvas image uploaded by a user, the framework can generate both preceding and succeeding states of the drawing process, and the generated states can be reused as inputs for further state generation. We observe that the user queries typically have one-to-one or many-to-many states, and in many cases, involve non-contiguous states. This necessitates a backend that solves a set-to-set problem with arbitrary combinations of past or future states. To this end, we repurpose video diffusion models to learn the set-to-set mapping with pretrained video priors. We implement the system with strong diffusion transformer backbones (*e.g.*, CogVideoX and LTXVideo) and high-quality data processing (*e.g.*, sampling short shots from long videos of real drawing records). Experiments show that the generated states are diverse in drawing contexts and resemble human drawing processes. This capability may aid artists in visualizing potential outcomes, generating creative inspirations, or refining existing workflows.

CCS Concepts: • **Applied computing** → **Fine arts**; **Media arts**.

Additional Key Words and Phrases: Digital painting, generative models, diffusion models

ACM Reference Format:
Lvmin Zhang, Chuan Yan, Yuwei Guo, Jinbo Xing, and Maneesh Agrawala. 2025. Generating Past and Future in Digital Painting Processes. *ACM Trans. Graph.* 44, 4 (August 2025), 13 pages. https://doi.org/10.1145/3731160

1 INTRODUCTION

Capturing, visualizing, and manipulating the editing process is fundamental and indispensable in digital creation and computer-aided

Authors' addresses: Lvmin Zhang, lvmin@cs.stanford.edu, Stanford University, USA; Chuan Yan, cyan3@gmu.edu, Stanford University, USA; Yuwei Guo, guoyw.nju@gmail.com, The Chinese University of Hong Kong, Hong Kong, China; Jinbo Xing, jbxing@cse.cuhk.edu.hk, The Chinese University of Hong Kong, Hong Kong, China; Maneesh Agrawala, maneesh@cs.stanford.edu, Stanford University, USA.

0730-0301/2025/8-ART $15.00
https://doi.org/10.1145/3731160

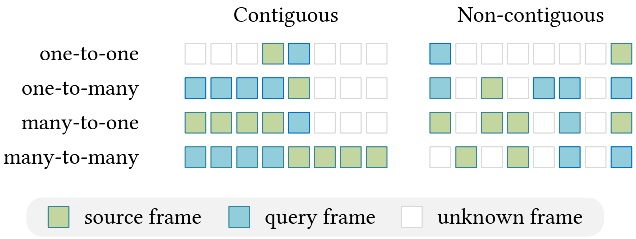


Fig. 2. **Various type of queries.** Different types of user queries may be requested by the frontend in various use cases. Either source or query can be one or multiple frames, and those frames can be non-contiguous in certain cases. For instance, querying a sketch from a finished drawing is a typical one-to-one non-contiguous query.

design. Decades of academic research and software development have contributed to various forms of process editing, like reproducing user operations (*e.g.*, undo/redo), recording rasterized canvases (*e.g.*, MS Paint), scriptable marcos (*e.g.*, MS Word's macro), visualizing processes (*e.g.*, Adobe Photoshop's history panel), parameterized history editing [Berthouzoz et al. 2011; Chen et al. 2016; Denning and Pellacini 2013; Denning et al. 2015; Grabler et al. 2009; Salvati et al. 2015], *etc.* Recording user edits as videos is another essential and versatile way for creators to explore process history, exchange ideas, share experiences, and study the details of decision making and design alternatives. Artists often consider design alternatives as an integral part of their process, *e.g.*, when creating a drawing, they may revisit earlier stages, re-imagine or redraw parts of their works to try different possible designs.

In the era of large generative models, can we further facilitate the process manipulation with large-scale pretrained models? Can we build a system capable of generating both the past and future of drawing process videos? Such a framework would not only allow users to explore possible outcomes but also inspire creative directions by visualizing alternative processes. This would provide new aspects of interactivity and aspiration, enabling creators to directly see the evolution of their work in reversive alternatives, reimagining earlier stages, or projecting potential future paths. From this perspective, how can we model the complicated relationships and interactions between different frames of a drawing process?

To answer these questions, we first observe that a system for generating past and future frames would need to respond to a range of queries, as shown in Fig. 2. To be specific, the queries may involve one-to-one or many-to-many mappings of drawing frames. For instance, users might wish to reconstruct the sketch from a finished artwork or visualize multiple future frames based on a partially completed piece. To design a smooth and efficient frontend, these backend queries may feature non-contiguous sections between frames or even more interwinded interactions. This complexity necessitates a robust backend capable of handling diverse frame combinations and dynamic frame structures while maintaining global coherence and contextual accuracy.

We propose to repurpose video diffusion models to learn such set-to-set mappings with pretrained video priors. Firstly, we introduce a partitioned 3D VAE, which allows for encoding contiguous and non-contiguous frame sequences while maintaining temporal coherence in the resulting latents. Secondly, we embed the operation steps (*i.e.*, the indices of the frames) into the temporal dimension of the neural hidden frames using causal projection layers, ensuring that the temporal layers match the step indices exactly. In this way, the learning can be stablized and the projection becomes robust in handling both contiguous and non-contiguous frame sections.

To this end, we train with strong diffusion transformer backbones (CogVideoX and LTXVideo) and implement high-quality data processing pipelines. Our data processing includes sampling short clips from long videos of real drawing records, preserving both contextual richness and variability. These strong backbones, paired with well-curated datasets, allow the framework to process diverse query types with robustness and coherency.

Experiments show that the framework handles diverse query types and combinations of frames, spanning a wide range of input styles and contexts, and generates outputs that resemble human drawing processes. These capabilities enable creative applications such as visualizing alternative process directions, exploring revisions and inspirations, or even generating new creating paths. Additionally, we provide ablative experiments to analyze the influence of each component.

In summary, (1) we motivate the problem of generating past and future frames of the drawing process and discuss the types of queries this system must handle; (2) we present the framework utilizing video diffusion models, enhanced with partitioned 3D VAEs and causal projection mechanisms, to address diverse query types effectively; (3) we provide practical applications and user-friendly interfaces for leveraging the framework in creative workflows, enabling users to explore alternative outcomes and visualize the generated evolutions of their work; (4) we perform extensive experiments, including ablative studies, to demonstrate the efficacy of each component and highlight the framework's robustness across various input styles and contexts; and (5) we explore additional applications, *e.g.*, creating unique special effects by generating different past/future processes from same input sequences.

2 RELATED WORK

2.1 Human Drawing Process

Human drawing is not just a sequence of strokes but a cognitive process involving observation, interpretation, and iteration. Early research explored human observation and perception, progressing to stroke-based rendering methods [Hertzmann 2003]. Stroke techniques for texture and tone emulation were also introduced [Salisbury et al. 1994]. Abstract image representations decomposing visual elements were proposed [Haeberli 1990]. Curved brush strokes of varying sizes enabled painterly rendering [Hertzmann 1998], while expressive brushwork mimicking impressionist art was addressed [Litwinowicz 1997].

Research has increasingly focused on mimicking the human painting process. Time-lapse videos capture artists' step-by-step approaches [Tan et al. 2015], while deep learning models replicate human-like strategies, such as generating vector path sequences based on images [Mo et al. 2021]. Models like SketchRNN [Ha and Eck 2018] and BézierSketch represent scalable, high-resolution stroke sequences [Das et al. 2020]. Neural painters optimize brushstrokes [Nakano 2019], and reinforcement learning models decompose images into strategic strokes [Huang et al. 2019]. Paint Transformer predicts strokes collectively, enhancing efficiency [Liu et al. 2021]. Stylized Neural Painting recreates global compositions with realistic textures [Zou et al. 2021], while IntelliPaint replicates human-like layering and semantic guidance [Singh et al. 2022]. Video models simulate querying the forward painting process through text description [Song et al. 2024] or the iterative artistic decisions based on a finished painting [Chen et al. 2024b; Zhao et al. 2020].

2.2 Digital Painting and Image Editing

Interactive painting methods now empower artists by combining human creativity with computational capabilities. Early methods addressed tasks like sketch cleanup, inking [Simo-Serra et al. 2018a,b], and grayscale colorization [Zhang et al. 2018, 2017]. With diffusion models, advanced tools have emerged for refining sketches into realistic images. Sketch-guided diffusion models provide control over image generation [Voynov et al. 2023], lightweight mapping networks enhance sketch realism [Roy et al. 2025], and abstraction-aware frameworks translating simple sketches into precise outputs [Koley et al. 2024]. EdgeGAN enables scene-level image creation from sketches [Gao et al. 2020], while spatially-adaptive normalization maintains photorealistic synthesis and semantic alignment [Park et al. 2019]. Additionally, multimodal conditioned image editing [Zhang et al. 2023] has enabled even greater flexibility in artistic expression and creative workflows. These methods can be grouped into various tasks: Conditioned content generation aligns images with text descriptions, sketches, or semantic maps [Sca 2024; Brooks et al. 2023; Xu et al. 2024]. Synthesizes photorealistic edits of coarsely modifications [Alzayer et al. 2024; Nitzan et al. 2024]. Transforming sketch-based editing into photorealistic images [Park et al. 2019; Qu et al. 2024]. Utilizing Diffusion-based approaches for image-to-image translation and refinement [Chen et al. 2024a; Pandey et al. 2024; Saharia et al. 2022]. Other approaches include realistic and exemplar-driven editing [Kawar et al. 2023; Yang et al. 2023; Zhang et al. 2018], user-guided interactive tools for intuitive modifications [Pan et al. 2023; Shi et al. 2024b], face and portrait-specific generation from rough inputs [Chen et al. 2020], and image harmonization techniques to ensure style and lighting consistency [Lu et al. 2023].

2.3 Video Diffusion and Generation

Video generation, transformed by diffusion models, has advanced from methods like VideoGAN [Vondrick et al. 2016] and video-to-video translation [Wang et al. 2018], which synthesized short clips using noise or segmentation masks, to video diffusion models [Ho et al. 2022] that achieve high-quality, temporally coherent outputs. Innovations such as Latent Diffusion Models [Rombach et al. 2022], text-to-video pipelines [Singer et al. 2023], and zero-shot animation with AnimateDiff [Guo et al. 2024] and ToonCrafter [Xing et al. 2024] adapt static image models for dynamic sequences. Classifier-free guidance [Ho and Salimans 2022] enables controllable video creation, while latent-space approaches [Blattmann et al. 2023] improve memory efficiency. Recent work focuses on scalability and motion consistency: spatiotemporal transformers [Menapace et al. 2024], unified editing frameworks [Bai et al. 2024], spatial-temporal diffusion for detailed motion [Bar-Tal et al. 2024], and explicit motion modeling [Shi et al. 2024a]. Additionally, PyramidFlow [Jin et al. 2024] introduces a pyramidal flow matching method that reinterprets the denoising trajectory as a series of pyramid stages, where only the final stage operates at the full resolution, so as to reduce computational costs. Tools like VideoCrafter [Chen et al. 2024c] leverage refined pretraining to achieve high-quality outputs with fewer constraints. Recent advancements in video generation have yielded open-source models [HaCohen et al. 2024; Weijie Kong and Jie Jiang 2024; Yang et al. 2024] rivaling closed-source counterparts.

3 METHOD

We build a framework that can generate the past and future drawing processes for digital paintings. Section 3.1 introduces the frame representation framework, defining how each step in a drawing is treated as an operation frame. Section 3.2 details the adaptation of video diffusion models, including strategies for handling temporal encoding and leveraging latent space compression. Section 3.3 presents the use of additional image diffusion models for efficient single-frame queries. Finally, Section 3.4 discusses implementation details and inference optimizations.

3.1 Framed Representation of Drawing Process

Considering a digital painting process with $s_{\max} \in \mathbb{Z}$ drawing steps, we denote the drawing canvas at the step $s \in [0, s_{\max}]$ as an RGB image $\boldsymbol{X}_s \in \mathbb{R}^{h \times w \times 3}$ with width w and height h. Intuitively, $\boldsymbol{X}_0$ is a pure blank image, and $\boldsymbol{X}_{s_{\max}}$ is the finished painting. In this paper, we call each step s an *operation step* and each canvas $\boldsymbol{X}_s$ a *frame*.

Aligning operation steps. An important representation in this framework is the use of a fixed number of max operation steps $s_{\max}$ (we use $s_{\max} = 1000$ by default). This ensures that, from the perspective of Positional Encoding (PE), the model can always be conditioned on a determined PE vector for all finished paintings. If the actual number of operation steps is smaller than 1000, we sample nearest neighbors; if the actual steps exceed 1000, we sample 1000 steps with randomized step skips as data augmentations (see also implementation details in Sec. 3.4).

Objective mapping. We establish backend models that can respond to different types of queries, *e.g.*, estimating one or many past/future frames from one or many existing frames. For all possible one-to-one or many-to-many queries, we can always formulate a set-to-set mapping $\mathcal{F}$ as

$$\mathcal{F} : \{\underbrace{\boldsymbol{X}_{s_1}, \ldots, \boldsymbol{X}_{s_n}}_{\text{source frames}}, \underbrace{s_1, \ldots, s_n}_{\text{source steps}}, \underbrace{q_1, \ldots, q_m}_{\text{queried steps}}\} \mapsto \{\underbrace{\boldsymbol{X}_{q_1}, \ldots, \boldsymbol{X}_{q_m}}_{\text{queried frames}}\}. \quad (1)$$

where we aim to estimate the queried frames $\{\boldsymbol{X}_{q_1}, \ldots, \boldsymbol{X}_{q_m}\}$ corresponding to the queried steps $\{q_1, \ldots, q_m\}$ based on the given source frames $\{\boldsymbol{X}_{s_1}, \ldots, \boldsymbol{X}_{s_n}\}$ and their respective source steps $\{s_1, \ldots, s_n\}$. To support various use cases, the source/queried steps (and the numbers m, n) are arbitrary.

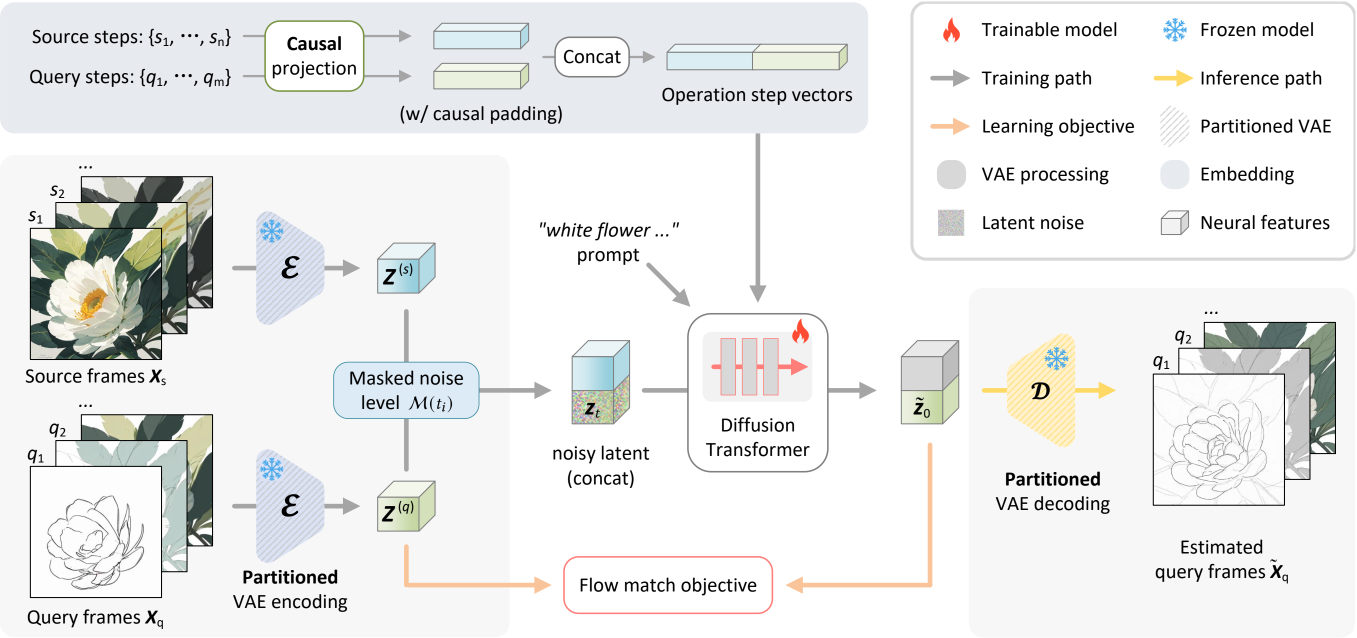


Fig. 3. **Framework overview.** We present an overview of the training framework. The diffusion transformer receive concatenated latents from sources and queries and only denoise the query part. The operation steps are embedded to the transformer with causal transforms to match the latent temporal dimension. Note that both source frames and query can be either multiple frames or one single frames.

3.2 Repurposing Video Diffusion Models

Preliminaries. Video diffusion models generate videos by learning to denoise data distribution (often in latent space) through a forward pass to gradually add noise to data and a reverse pass to reconstruct the data. We here discuss latest architectures of Diffusion Transformers (DiTs) and flow match scheduling, whereas other architectures can be adapted with similar formulations. Typical base models in this category are CogVideoX [Yang et al. 2024], HunyuanVideo [Weijie Kong and Jie Jiang 2024], and LTXVideo (LTXV) [HaCohen et al. 2024]. Rectified-flow models map noisy latents

$$z_{t_i} = (1 - t_i)z_0 + t_i\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

to clean latents, z_0 where $t_i \in (0, 1]$ is the diffusion timestep. Most recent video diffusion models use latent diffusion with VAEs to compress data with

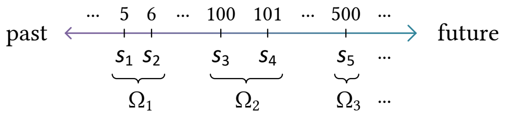
$$z_0 = \mathcal{E}(\boldsymbol{X}), \quad \hat{\boldsymbol{X}} = \mathcal{D}(z_0), \tag{3}$$

where $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$ are encoder and decoder of the VAE, respectively, *e.g.* with LTXV, this process compress image sequence $\boldsymbol{X} \in \mathbb{R}^{f \times h \times w \times 3}$ into $z_0 = \mathcal{E}(\boldsymbol{X}) \in \mathbb{R}^{\frac{f}{8} \times \frac{h}{32} \times \frac{w}{32} \times 128}$ with f, h, w being frame count, width, and height. To learn the DiT generator $G_\theta(\cdot)$, the learning objective is

$$\mathbb{E}_{z_0, \boldsymbol{c}, t_i \sim \mathcal{L}(0,1), \boldsymbol{\epsilon} \sim \mathcal{N}(0,1)} \left\| (\boldsymbol{\epsilon} - z_0) - G_\theta(z_{t_i}, t_i, \boldsymbol{c}) \right\|_2^2, \tag{4}$$

where $\boldsymbol{c}$ is a set of conditions like text prompts, and $t \sim \mathcal{L}(0, 1)$ is the shifted logit-normal distribution [Esser et al. 2024] for flow match timestep sampling.

Partitioned 3D VAE. We encode a set of drawing frames $\boldsymbol{X}_{s_{1\ldots n}}$ with the pretrained VAE encoder $\mathcal{E}(\cdot)$. Video diffusion frameworks typically use 3D VAEs with causal convolutions to encode multiple frames together. This encoding allows for temporal coherency in reconstructing multiple frames, but would introduce unnecessary ghosting artifacts when encoding non-contiguous frames. When frame indices $s_{1\ldots n}$ contain non-contiguous sub-sequences, we propose to partition them into contiguous sections, *e.g.*,



and then encode them independently. Considering $\Omega_{1\ldots k}$ representing the partitioned contiguous sub-sequences of frame indices, each section $\boldsymbol{X}_{\Omega_i}$ can be encoded into the latent $\mathcal{E}(\boldsymbol{X}_{\Omega_i})$, and the final encoded result are concatenated

$$z = \left[\mathcal{E}(\boldsymbol{X}_{\Omega_1}) \ldots \mathcal{E}(\boldsymbol{X}_{\Omega_k})\right] \quad , \tag{5}$$

and in this process, one special case is that a section Ω_i contains only one frame (or less than the minimal frame number specified by the VAE causal convolutions). In this case the inputs are padded by the causal convolution, *e.g.* with LTXV, n frames are padded with duplicated prefixes and then encoded to $\lfloor 1 + \frac{n-1}{8} \rfloor$ latents.

Conditioning operations steps. Considering that each drawing frame $\boldsymbol{X}_s$ is paired with a operation step index s, when multiple frames $\boldsymbol{X}_{s_{1\ldots n}}$ are encoded by 3D VAE into a latent cube, the indices $s_{1\ldots n}$ need to be projected to match the latent temporal dimension. As shown in Fig. 4, we propose to project the operation step indices

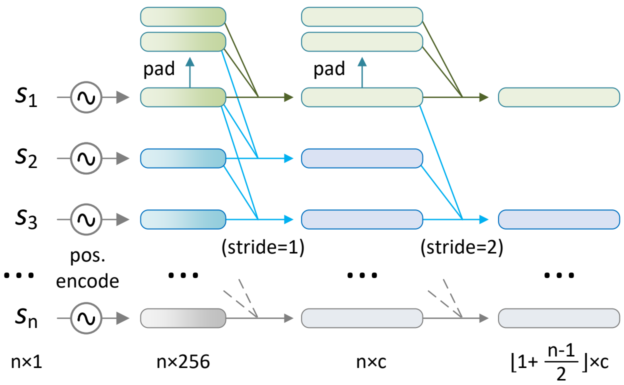


Fig. 4. **Causal padding for operations step conditioning.** We illustrate the 2D convolution layers for projecting the step feature vectors to the latent shape that matches the latent temporal dimension. For layers with any kernel size $k_s \in Z$, the causal padding always pad the first element $k_s - 1$ times. No matter how large the number n is, the first vector always only encode features from the first input. In this example, we use 256 frequencies for positional encoding, and c means hidden state channels of one layer.

with a set of causal 2D convolution layers that are padded with exactly the same configurations with the 3D VAE, *e.g.* with LTXV, n indices will be encoded into $\lfloor 1 + \frac{n-1}{8} \rfloor$ vectors to match the latents (The layers in Fig. 4 will compress latent temporal dimension by 2, and we repeat this 3 times). The newly initialized projection layers are then gated by a zero-initialized linear layer, and added to the timestep embedding layers in DiTs transformer models.

Source and query frames. We condition the model on source (existing) frames and generate queried (unknown) frames. To achieve this conditioning, we always view source and query frames as non-contiguous sections in the aforementioned partitioned VAE encoder. In other words, source and query frames always have independent latent temporal entries. For instance, considering n contiguous source frames and m contiguous query frames, we always encode them into $\lfloor 1 + \frac{n-1}{8} \rfloor + \lfloor 1 + \frac{m-1}{8} \rfloor$ latents (*i.e.*, not $\lfloor 1 + \frac{n+m-1}{8} \rfloor$ even if those query frames are next to source frames). In this way, we can set the diffusion timestep (diffusion noise level) t_i to zero for the source frames during training, so that the diffusion model will only denoise the queried frames and use the source frames as references. We denote this transform by masking $\mathcal{M}(t_i)$ to zero-out the source frame part in diffusion time steps t_i.

Learning objective. We have the joint objective in form of Eq. 4 as

$$\mathbb{E}_{z,c,s,t_i \sim \mathcal{L}(0,1), \epsilon \sim \mathcal{N}(0,1)} \left\| (\epsilon - z) - G_\theta \left(z_{\mathcal{M}(t_i)}, \mathcal{M}(t_i), c, s \right) \right\|_2^2, \quad (6)$$

with c being model-specific extra conditions, *e.g.* text prompts for LTXV, CogVid, *etc.* The clean latents z are encoded by the partitioned 3D VAE with Eq. 5.

3.3 Additional Model Variants and Efficient Inference

Though the video diffusion model itself can already process all types of user queries, the experience of many typical queries can be improved by training additional model variants to build practical tools and responsive interfaces. For instance, a special yet highly common type of user input is a case where both the source and query only specify a single frame, *e.g.*, directly querying the finished drawing from a sketch, or querying a rough sketch of a finished drawing, *etc.* One can achieve higher visual quality and reduced computation overheads by training a dedicated image diffusion model. Building more model variants also allows for wider visual effects, *e.g.*, to benefit from community LoRAs for specific models, *etc.* The advantages of image-based diffusion models are also discussed in related works like InversePainting [Chen et al. 2024b].

Additional image diffusion variant. To be specific, we train an independent image diffusion model SDXL [Podell et al. 2023] for queries that have a single source and queried frame. We add 4 zero-initialized channels to SDXL input convolution projection to receive the latents of source X_s, and extend the timestep embedding projection to receive two extra scalars: the source step s and query step q. The model is then trained using SDXL's diffusion objective.

Efficient inference scheduling. With multiple available models, the inference efficiency can be tweaked on different deployment devices. We provide a default scheduling: (1) if both the source and query are single frames, respond with the additional image diffusion model; (2) if the source is a single frame and the query is a large enough set of frames (> 500), we first use the additional image model to generate one intermediate frame for every interval of 200 steps, and then use the video diffusion model to generate the remaining frames in every interval, with the frames from the image diffusion model as extra inputs; (3) for all remaining types of user queries, only respond with the video diffusion model. The performance differences between model variants will be discussed in ablation experiments.

3.4 Data and Implementation Details

Data preparation. We start with a progressive collection of long videos recording the drawing processes. These videos were processed using a pipeline to sample a much larger set of video clips, with durations aligned to the context length supported by pretrained video models. The data collection happened during winter 2018 to autumn 2020. We reached out to artists one-by-one and obtained consents from about 19 artists. At the time of collection, more than half of the participants were students enrolled in an art school class, others were professional artists with careers related to art, and the occupations of the remaining participants were unknown. Among the students, more than half were highly skilled and capable of independently creating commercial artworks, while the remaining students were beginners, proficient only in sketching or replicating existing works. The drawing process videos were recorded using various software. In total, we collected many long videos, with each video lasting for hours. The content distribution covers diverse subjects including humans (various ages and genders), non-humans (animals, robots, *etc.*), scenes (outdoor, indoor), plants, toys, and fantasy elements. Stylistically, the majority comes from commercial digital paintings (concept art, promotional illustrations, *etc.*) with other styles like watercolor, doodle, surrealism, *etc.* We detected abrupt changes in those long videos with LTX's criteria [HaCohen et al. 2024] to obtain shots, and then filtered shots with insignificant

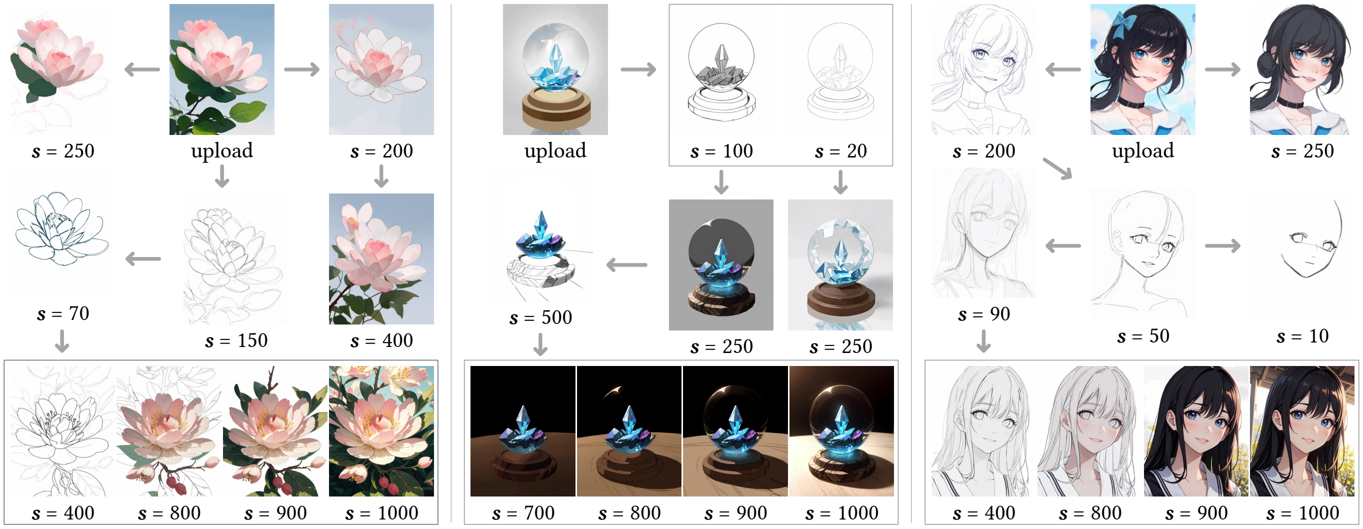


Fig. 5. **Examples of user explorations.** We show that users can upload canvas images, query past or future frames of the drawing process. The images marked with "upload" are user inputs, while all others are outputs. Images in grey rectangle are generated in one same inference. Model input prompts are captioned by WD14 tagger from input images.

changes. We view 145 frames as the default context length number. For shots longer than 145 frames, we use a randomized *step skip* — we separate the shot into 145 intervals and randomly sample one frame from each interval. The randomized step skip sampling is performed at most 32 times for each shot longer than 145 frames. In this way, each shot becomes a video clip sample aligned with the video models' context length. In total, we obtained the final dataset of 20k video clips.

Training Details. We use the Adafactor [Shazeer and Stern 2018] optimizer with a learning rate of 1e-5 to train the diffusion transformers in bf16 precision. We train two video diffusion models with LTXVideo and CogVideoX 1.5. We also train an additional image model with SDXL (Sec. 3.3). The training devices are 8x H100 80G. The LTXVideo is trained for about 4 days. The SDXL image model is trained for about 3 days. The CogVideoX 1.5 model is trained for about 9 days. We modify the context length of CogVideoX 1.5 to 145 frames. For LTXVideo and CogVideoX, we use buckets of 512px resolution. For SDXL image training, we use 1024px resolution buckets. To maximize the batch size, we always enable diffusers' gradient check-pointing in all transformer blocks. To stabilize training, we set max gradient clip norm to 0.5. We do not use any EMA weights. Unless otherwise mentioned, for all inference and training, we always use WD14 tagger [SmilingWolf 2022] to get prompts from the input frame image with the largest step index.

4 EXPERIMENTS

4.1 Qualitative Results

Interactive exploration. As shown in Fig. 5, our method allows users to upload canvas images and explore the drawing process by moving forward or backward through different steps. This exploration results in various intermediate canvases. When users find a canvas that aligns with their vision, they can continue generating steps forward or backward to further refine or visualize alternative outcomes. The intermediate canvases generated during this process can also be used for other creative purposes.

Table 1. **Quantitative tests.** We test the reconstruction metrics using different methods and our ablative architectures. Video methods are measured with mean metrics over all frames, while image models are measured as the mean value of all images.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
[Zhao et al. 2020]	12.27	0.4847	0.6851
[Song et al. 2024]	13.52	0.4901	0.6143
Ours (LTXVideo)	16.31	0.6510	0.4259
Ours (CogVideo)	17.04	0.6712	0.4022
Ours (image model, SDXL)	15.98	0.5995	0.4515

Generating drawing processes. As shown in Fig. 6, our method can accept one or multiple user-input canvas images and generate frames representing either the past, future, or intermediate states of the drawing process. We also observe that our method generalizes well across diverse content types, including plants, animals, landscapes, portraits, and food. This versatility demonstrates the robustness and flexibility of the framework in handling a wide range of artistic subjects.

Generating multiple drawing processes. As shown in Fig. 7, our method can generate different drawing sequences from the same user input using various random seeds. This allows for the creation of multiple possible outcomes, each with unique artistic decisions. For example, as illustrated in the first row, a sketch of an animal can evolve into either a rabbit or a dog, depending on the input prompt.

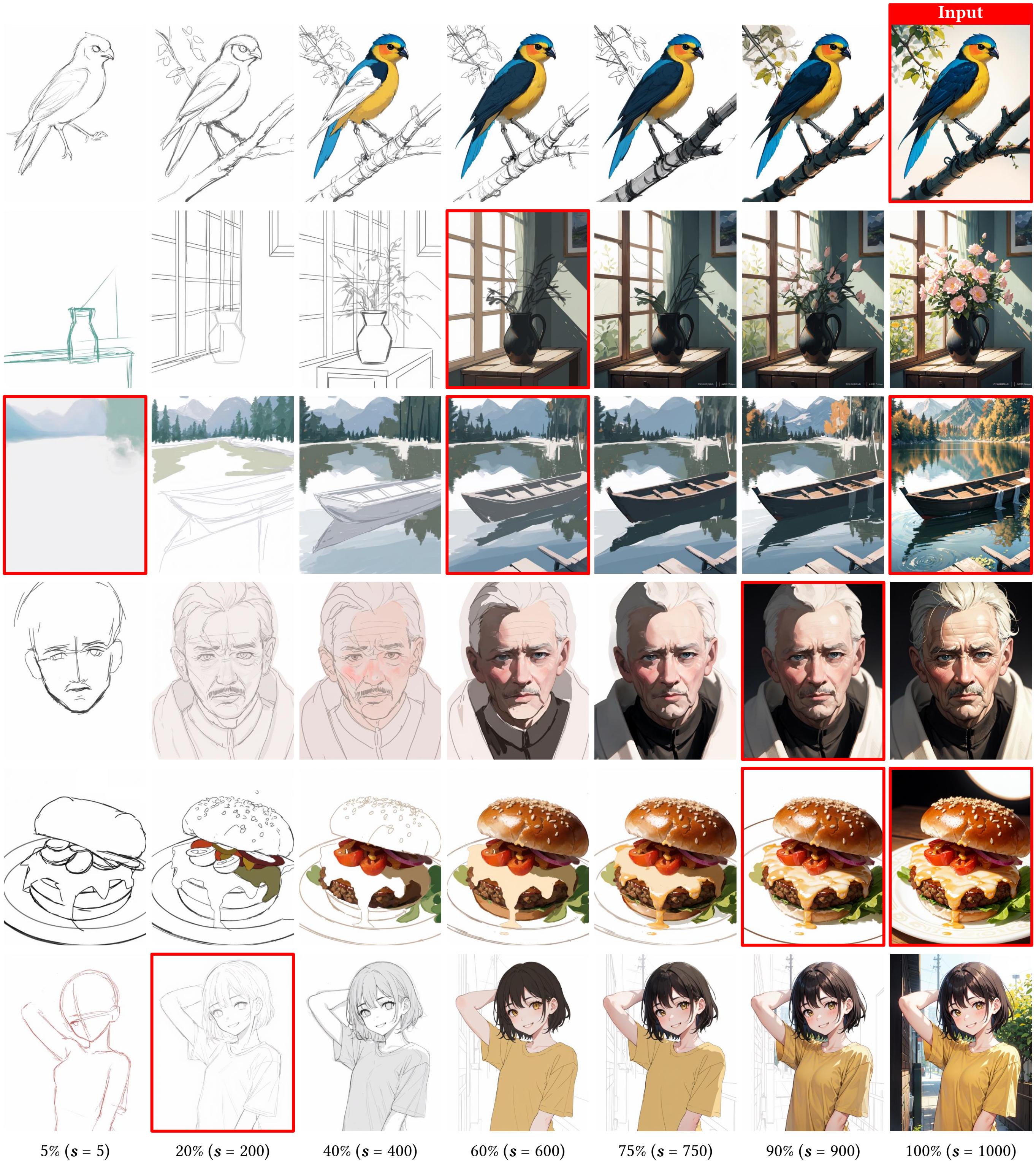


Fig. 6. **Qualitative results of process generation with single random seed.** Images in red rectangles are inputs while all others are outputs. The second, forth, sixth rows uses prompts “masterpiece, a pot of flower in room”, “masterpiece, a portrait of a handsome man”, “masterpiece, best quality, 1girl, shirt, outdoor”. All other rows uses WD14 tagger prompts.

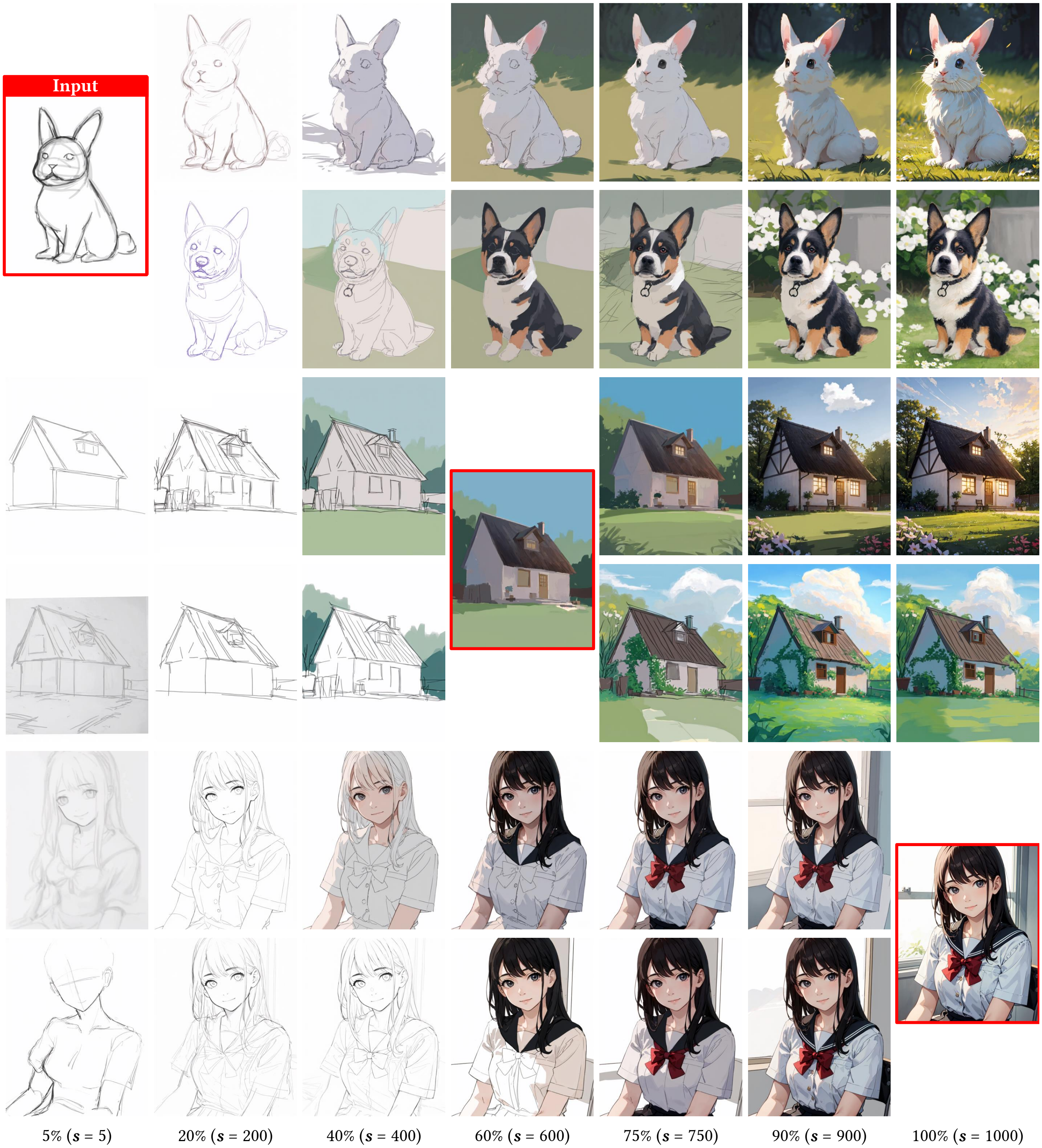


Fig. 7. **Qualitative results of process generation with multiple random seeds.** Images in red rectangles are inputs while all others are outputs. The first and second rows uses prompts "masterpiece, a artwork of a bunny, outdoor", "masterpiece, a artwork of a dog, outdoor". All other rows uses WD14 tagger prompts.

Table 2. **User Study.** We conduct user study with two objectives: to study which approach produces results with highest quality, and to study which approach is most practical for editing, offering better responsiveness (*e.g.*, faster speed). Previous methods are not involved in responsiveness test since they are not interactive applications. Numbers are mean preference rate with best in bold.

Candidate	Quality ↑	Responsiveness ↑
[Zhao et al. 2020]	1.20±0.0%	/
[Song et al. 2024]	3.61±1.21%	/
Ours (LTXVideo)	18.07±4.3%	27.49±9.7%
Ours (CogVideo)	20.48±5.2%	8.46±6.5%
Ours (LTXVideo + image model)	26.51±6.1%	**47.43±9.1%**
Ours (CogVideo + image model)	**30.12±5.8%**	16.62±5.2%

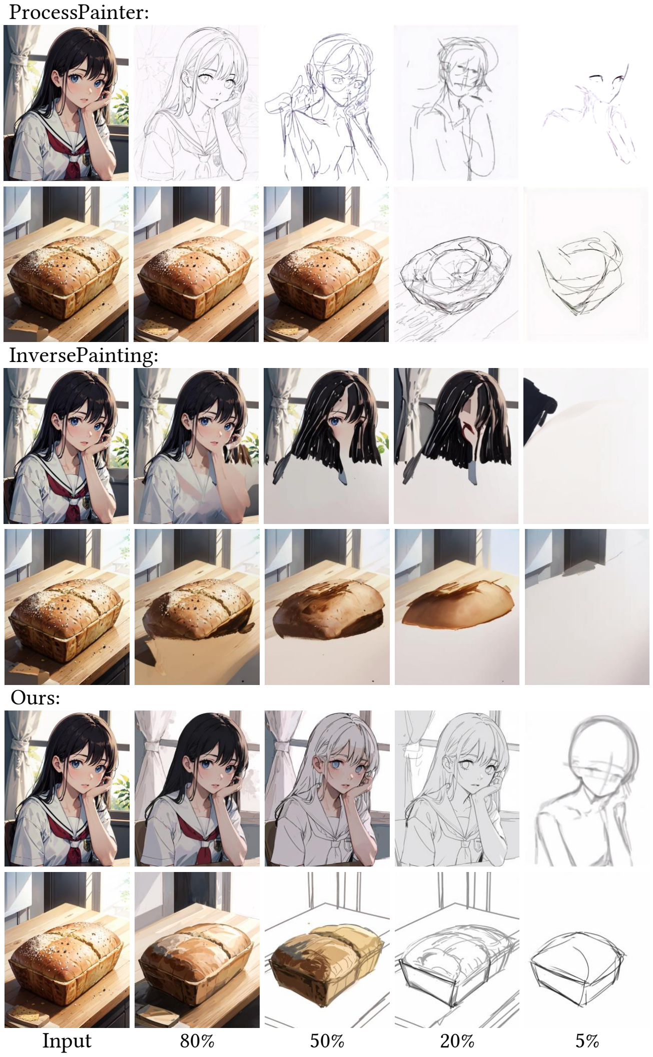


Fig. 8. **Difference to previous methods.** We present the difference between our approach and previous methods: ProcessPainter [Song et al. 2024] and InversePainting [Chen et al. 2024b]. Note that ProcessPainter suggests finetuning the model on target datasets, and we conducted the finetune on our same dataset.

4.2 Visual Comparison

As shown in Fig. 8, we present a comparison between our approach and previous methods, including ProcessPainter [Song et al. 2024] and InversePainting [Chen et al. 2024b]. Due to the lighter base models employed by ProcessPainter, it struggles to fit significant dynamic changes, such as transitions between sketching, outlining, and coloring stages. On the other hand, InversePainting models the drawing process as a segmentation-like task, which makes it unsuitable for handling drawing sequences that involves sketching behaviors. Our approach based on stronger foundational models and trained on high-quality aligned data, generates high-quality frames and coherent drawing processes.

4.3 Quantitative Results

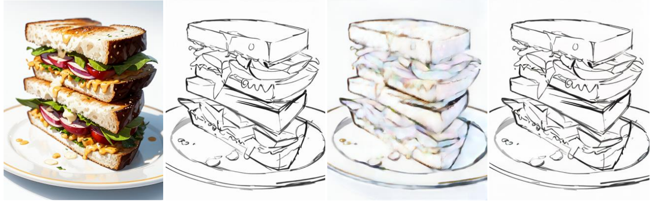
We evaluate our method on a held-out 5% subset of the training data, which the model has not seen during the training. We test different approaches, including three of our model variants and two of prior methods [Song et al. 2024; Zhao et al. 2020]. Quantitative metrics such as PSNR, SSIM, and LPIPS are used for evaluation, as shown in Table 1. Our method, particularly when using the CogVideo backbone, achieves the best results across all three metrics. We also test alternative configurations of our model, such as using LTXVideo as the base model and a standalone image diffusion model. These variants also outperform the prior methods in terms of reconstruction quality. It is worth noting that InversePainting [Chen et al. 2024b] is excluded from this comparison due to its reliance on a segmentation-based methodology, which is incompatible with our dataset.

4.4 User Study

We conducted a user study to evaluate both the output quality and the practicality of our method. Specifically, we compared various configurations of our framework against two prior methods [Song et al. 2024; Zhao et al. 2020]. Due to the significant structural and output differences of InversePainting [Chen et al. 2024b], it was excluded in the evaluation.

Setup. We selected 50 in-the-wild digital drawings that were unseen by the model during training. Participants were invited to modify these drawings using our framework to create 50 corresponding sequences of drawing processes. In addition to evaluating the generated sequences, we collected feedback on the usability of the different tools. Each participant used every configuration of our method at least once. To ensure fairness, prior methods were applied to the same input data after the initial user feedback collection for comparison.

Participants. A total of 13 individuals participated in the study, comprising one student, eight crowd-sourced workers recruited from online platforms, and four experienced associate artists with prior drawing expertise.



Frame A (end of prv. section) | Frame B (start of next section) | Recon. B (w/o partitioned VAE) | Recon. B (proposed)

Fig. 9. **Influence of Partitioned VAE.** We show two connected sections with non-contiguous frames and the VAE reconstruction of second frame. Without the Partitioned VAE design, the second frame will have blurred reconstruction caused by latent temporal compression.

Evaluating Quality. To assess the output quality, participants were asked to identify the drawing sequence they considered to have the highest quality for each task. The method with the highest average preference rate was considered the best in terms of quality.

Evaluating Responsiveness. Responsiveness was evaluated as an advanced metric of system speed and interactivity. Beyond measuring absolute processing time for generating frames, responsiveness accounts for the user's experience during iterative exploration. For example, participants could opt for multiple smaller generations (e.g., generating one frame at a time) rather than a single large batch, leading to a more fluid and interactive experience.

Metric. We used the average user preference rate as the metric. For each input, we calculate the percentage of participants who preferred a specific method. This process was repeated over five rounds, and the results were averaged to calculate the mean preference rate, with standard deviation used to indicate measurement error.

Results. As shown in Table 2, our method using the CogVideo backbone achieved the highest perceived quality scores, though its slower processing speed was noted. In terms of responsiveness, users preferred our variant with the LTXVideo backbone, which achieved faster inference times. Moreover, all configurations that utilized the additional image model ("+ image model") for scheduling (Section 3.3) outperformed others in both quality and responsiveness. These results highlight the advantages of our framework, which combines strong base models with effective training strategies, over prior methods.

4.5 Ablative Study

Partitioned VAE. As shown in Fig. 9, we evaluate the impact of adopting the Partitioned VAE design. Without this design, encoding non-contiguous frames introduces artifacts during the reconstruction process, such as blurring or ghosting caused by temporal compression in the latent space. By using the Partitioned VAE, non-contiguous sections are encoded independently, resulting in clean reconstructions and improved visual quality, especially when handling complex queries involving non-contiguous frame sequences.

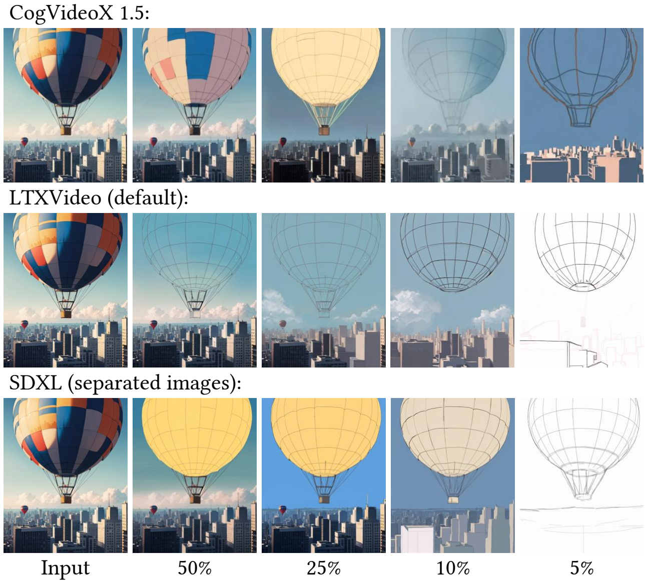


Fig. 10. **Visualization of ablative backends.** Different backends yield different visual results. The image model SDXL is trained to predict separate frames while other video models estimate frames jointly.

Table 3. **Runtime comparison of different model backends.** The reported times represent baseline performance on a Nvidia L40S GPU without advanced optimization techniques. Further speedups are possible using acceleration methods such as TeaCache. All measurements are in seconds.

Model	1 frame	65 frames	129 frames
CogVideoX 1.5 [Yang et al. 2024]	7.2s	193.1s	327.4s
LTXVideo [HaCohen et al. 2024]	4.8s	8.4s	41.3s
SDXL (image model)	3.9s	251.2s	495.1s

Table 4. **Abrupt transitions in drawing processes.** We analyze the frequency of different types of abrupt transitions per artwork in both human drawing recordings and our generated outputs. These transitions reflect common artistic behaviors that may appear as inconsistencies in the drawing process. Numbers represent average count per artwork in 50 samples. The numbers are counted by humans. Note that the operations are counted from the record videos and does not involve any meta data of the underlying operations (which does not exist for AI-generated records).

Abrupt Transition Types	Human	Proposed
Element visibility change	3.2±0.4	2.8±0.5
Element resizing and warping	3.7±0.3	2.1±0.4
Element import or delete	2.4±0.5	2.2±0.3
Color curves adjustment	1.5±0.2	1.8±0.3
Layer order change	2.3±0.4	2.5±0.4
Total	13.1±0.7	11.4±0.8

Different backend models. As shown in Fig. 10 and Table 3, we compare different model backends for our framework. While SDXL (image model) is efficient for single-frame generation (3.9s), LTXVideo

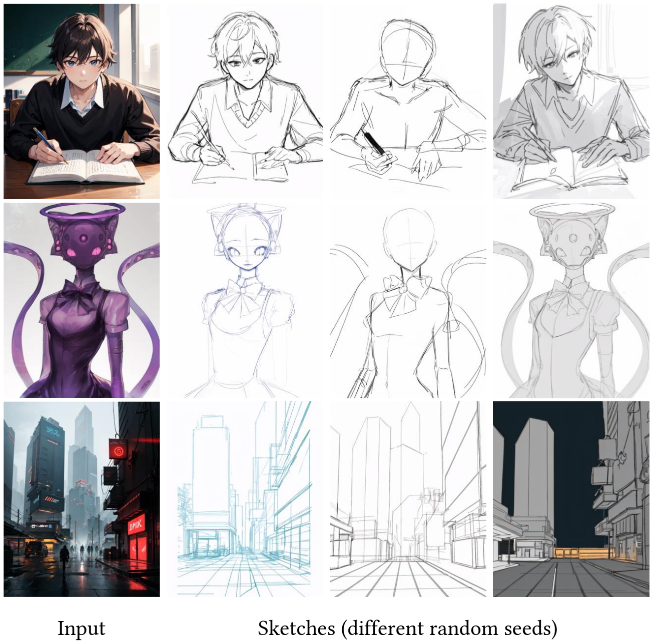


Fig. 11. **Additional application: sketch generation.** By setting the step to a low value (*e.g.*, $s = 50$ in this figure) and changing random seeds, the framework can generate sketches with different artistic decisions.

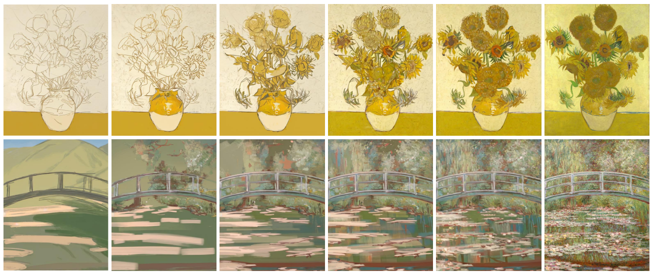


Fig. 12. **Results of non-digital paintings.** We present artworks from Van Gogh and Claude Monet. ©public domain.

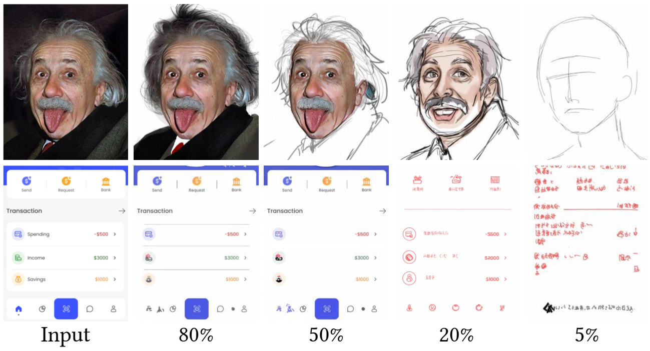


Fig. 13. **Limitation.** When processing images with styles too different to pretrained paintings, the process may have style offsets. When processing out-of-scope images like UI designs, errors will occur with text and structure.

demonstrates superior efficiency for multi-frame scenarios (0.13s per frame for 65 frames). CogVideoX 1.5 produces similar visual quality but at significantly higher computational cost. Video models capture essential temporal patterns in drawing processes. LTXVideo strikes a balance between processing time and quality, making it more suitable for interactive applications. Additionally, incorporating the SDXL image model for specific query types (*e.g.*, single-frame queries) enhances frontend responsiveness and user experience. This demonstrates that a hybrid approach leveraging both video and image models can optimize performance for different query scenarios.

4.6 Additional Discussions

Sketch generator. As shown in Fig. 11, our method generate diverse sketches by setting the target step to a lower value (e.g., $s = 50$) and varying the random seed. Unlike traditional edge-detection methods, these sketches capture artistic decisions and structural abstractions learned from the dataset, reflecting a natural drawing process. These generated sketches can serve as valuable resources for training other models, such as sketch-to-image generators, or for inspiring artists during the initial stages of the creative process.

Generalization to additional art styles. As shown in Fig. 12, we explore how our framework handles classical artworks from Van Gogh and Monet that differ from the digital paintings in our training dataset. The underlying transformer backbone of the model facilitates the transfer and generalization to these different artistic styles. When reconstructing drawing processes for these works, we observe that the model follows a pattern of initially composing color blocks and later adding finer brush strokes, similar to how a digital artist might approach recreating such non-digital paintings.

Abrupt transitions and human behaviors. Table 4 quantifies the frequency of different types of abrupt transitions in both human and AI-generated drawing processes. These transitions reflect common artistic behaviors that might appear as inconsistencies but are actually essential to the real drawing processes. Human artists frequently toggle layer visibility, resize and warp elements, import or delete content, adjust color curves, and change layer ordering. Our model captures these behaviors with similar frequencies, demonstrating its ability to learn the procedural aspects of artistic creation.

4.7 Limitation

As shown in Fig. 13, our method encounters limitations when processing input images that significantly differ in style from the training data. For instance, inputs such as real photographs or UI design layouts can result in style inconsistencies or artifacts, including errors in text rendering or structural details. While this issue may be alleviated by expanding the dataset to include a broader range of styles, certain tasks, such as reconstructing drawing processes for real-world photographs, remain inherently ill-conditioned problems. In these cases, an optimal mapping distribution may not exist.

5 CONCLUSION

We introduce a framework for generating past and future states of digital painting processes, enabling artists to explore, refine, and visualize creative process videos or individual images. The framework is based on latest video diffusion models with strong transformer backbones. We propose partitioned 3D VAEs for handling non-contiguous frame sequences, and robust data processing pipelines to address the complexities of the set-to-set mappings across diverse input queries. Extensive experiments and user studies demonstrate that our framework produces high-quality, coherent drawing processes while offering interactive responsiveness. The resulting model can reconstruct drawing sequences that closely ensemble human process and generating alternative artistic directions. We discuss various ways to make use of this system and demonstrate creative manipulations with applications such as sketch generation, re-imagining early stages, or visualizing multiple potential outcomes from a single input.

ACKNOWLEDGMENTS

This work was partially supported by the Brown Institute for Media Innovation, by Google through their affiliation with Stanford Institute for Human-centered Artificial Intelligence (HAI) and by a Hoffman-Yee HAI grant.

REFERENCES

2024. Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport. In *The Thirteenth International Conference on Learning Representations.*

Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. 2024. Magic Fixup: Streamlining Photo Editing by Watching Dynamic Videos. arXiv:2403.13044 [cs.CV] https://arxiv.org/abs/2403.13044

Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. 2024. UniEdit: A Unified Tuning-Free Framework for Video Motion and Appearance Editing. *arXiv preprint arXiv:2402.13185* (2024).

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. 2024. Lumiere: A Space-Time Diffusion Model for Video Generation. In *SIGGRAPH Asia 2024 Conference Papers (SA '24).* Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3680528.3687614

Floraine Berthouzoz, Wilmot Li, Mira Dontcheva, and Maneesh Agrawala. 2011. A Framework for content-adaptive photo manipulation macros: Application to face, landscape, and global manipulations. *ACM Trans. Graph.* 30, 5, Article 120 (Oct. 2011), 14 pages. doi:10.1145/2019627.2019639

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

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Vancouver, BC, Canada, 18392–18402. doi:10.1109/CVPR52729.2023.01764

Bowei Chen, Yifan Wang, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. 2024b. Inverse Painting: Reconstructing The Painting Process. In *SIGGRAPH Asia 2024 Conference Papers (SA '24).* Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3680528.3687574

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024c. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7310–7320.

Hsiang-Ting Chen, Li-Yi Wei, Björn Hartmann, and Maneesh Agrawala. 2016. Data-driven adaptive history for image editing. In *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games* (Redmond, Washington) *(I3D '16).* Association for Computing Machinery, New York, NY, USA, 9 pages. doi:10.1145/2856400.2856417

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2024a. TextDiffuser: Diffusion Models as Text Painters. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23).* Curran Associates Inc., Red Hook, NY, USA, 9353–9387.

Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. 2020. DeepFaceDrawing: Deep Generation of Face Images from Sketches. *ACM Trans. Graph.* 39, 4 (Aug. 2020), 72:72:1–72:72:16. doi:10.1145/3386569.3392386

Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. 2020. Béziersketch: A generative model for scalable vector sketches. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16.* Springer, 632–647.

Jonathan D. Denning and Fabio Pellacini. 2013. MeshGit: diffing and merging meshes for polygonal modeling. *ACM Trans. Graph.* 32, 4, Article 35 (July 2013), 10 pages. doi:10.1145/2461912.2461942

Jonathan D. Denning, Valentina Tibaldo, and Fabio Pellacini. 2015. 3DFlow: continuous summarization of mesh editing workflows. *ACM Trans. Graph.* 34, 4, Article 140 (July 2015), 9 pages. doi:10.1145/2766936

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206 [cs.CV]

Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. 2020. SketchyCOCO: Image Generation From Freehand Scene Sketches. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 5173–5182. doi:10.1109/CVPR42600.2020.00522

Floraine Grabler, Maneesh Agrawala, Wilmot Li, Mira Dontcheva, and Takeo Igarashi. 2009. Generating photo manipulation tutorials by demonstration. *ACM Trans. Graph.* 28, 3, Article 66 (July 2009), 9 pages. doi:10.1145/1531326.1531372

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations* (2024).

D. Ha and D. Eck. 2018. A neural representation of sketch drawings. In *International Conference on Learning Representations (ICLR).*

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. 2024. LTX-Video: Realtime Video Latent Diffusion. arXiv:2501.00103 [cs.CV] https://arxiv.org/abs/2501.00103

Paul Haeberli. 1990. Paint by numbers: abstract image representations. *SIGGRAPH Comput. Graph.* 24, 4 (Sept. 1990), 8 pages. doi:10.1145/97880.97902

Aaron Hertzmann. 1998. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98).* Association for Computing Machinery, New York, NY, USA, 8 pages. doi:10.1145/280814.280951

A. Hertzmann. 2003. A survey of stroke-based rendering. *IEEE Computer Graphics and Applications* 23, 4 (2003), 70–81. doi:10.1109/MCG.2003.1210867

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *arXiv:2204.03458* (2022).

Zhewei Huang, Shuchang Zhou, and Wen Heng. 2019. Learning to Paint With Model-Based Deep Reinforcement Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* IEEE, Seoul, Korea (South), 8708–8717. doi:10.1109/ICCV.2019.00880

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. 2024. Pyramidal Flow Matching for Efficient Video Generative Modeling. arXiv:2410.05954 [cs.CV]

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Vancouver, BC, Canada, 6007–6017. doi:10.1109/CVPR52729.2023.00582

Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2024. It's All About Your Sketch: Democratising Sketch Control in Diffusion Models. In *CVPR.*

Peter Litwinowicz. 1997. Processing images and video for an impressionist effect. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97).* ACM Press/Addison-Wesley Publishing Co., USA, 8 pages. doi:10.1145/258734.258893

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. 2021. Paint Transformer: Feed Forward Neural Painting with Stroke Prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* IEEE, Montreal, QC, Canada, 6578–6587. doi:10.1109/ICCV48922.2021.00653

Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing Zhang. 2023. Painterly Image Harmonization Using Diffusion Model. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 233–241. doi:10.1145/3581783.3612451

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, and Sergey Tulyakov. 2024. Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 7038–7048. doi:10.1109/CVPR52733.2024.00672

Haoran Mo, Edgar Simo-Serra, Chengying Gao, Changqing Zou, and Ruomei Wang. 2021. General Virtual Sketching Framework for Vector Line Art. *ACM Transactions on Graphics* 40, 4 (July 2021), 51:1–51:14. doi:10.1145/3450626.3459833

Reiichiro Nakano. 2019. Neural Painters: A learned differentiable constraint for generating brushstroke paintings. *ArXiv* abs/1904.08410 (2019). https://api.semanticscholar.org/CorpusID:120367960

Yotam Nitzan, Zongze Wu, Richard Zhang, Eli Shechtman, Daniel Cohen-Or, Taesung Park, and Michaël Gharbi. 2024. Lazy Diffusion Transformer for Interactive Image Editing. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXIV* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 55–72.

Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. In *ACM SIGGRAPH 2023 Conference Proceedings (SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3588432.3591500

Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. 2024. Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7695–7704. doi:10.1109/CVPR52733.2024.00735

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2332–2341. doi:10.1109/CVPR.2019.00244

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

Linzi Qu, Jiaxiang Shang, Hui Ye, Xiaoguang Han, and Hongbo Fu. 2024. Sketch2Human: Deep Human Generation with Disentangled Geometry and Appearance Constraints. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–14. doi:10.1109/TVCG.2024.3403160

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 10674–10685. doi:10.1109/CVPR52688.2022.01042

Prasun Roy, Saumik Bhattacharya, Subhankar Ghosh, Umapada Pal, and Michael Blumenstein. 2025. D-Sketch: Improving Visual Fidelity of Sketch-to-Image Translation with Pretrained Latent Diffusion Models without Retraining. In *Pattern Recognition*, Apostolos Antonacopoulos, Subhasis Chaudhuri, Rama Chellappa, Cheng-Lin Liu, Saumik Bhattacharya, and Umapada Pal (Eds.). Springer Nature Switzerland, Cham, 277–292. doi:10.1007/978-3-031-78389-0_19

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-Image Diffusion Models. In *ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3528233.3530757

Michael P. Salisbury, Sean E. Anderson, Ronen Barzel, and David H. Salesin. 1994. Interactive Pen-and-Ink Illustration. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '94)*. Association for Computing Machinery, New York, NY, USA, 101–108. doi:10.1145/192161.192185

Gabriele Salvati, Christian Santoni, Valentina Tibaldo, and Fabio Pellacini. 2015. MeshHisto: collaborative modeling by sharing and retargeting editing histories. *ACM Trans. Graph.* 34, 6, Article 205 (Nov. 2015), 10 pages. doi:10.1145/2816795.2818110

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 4596–4604.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. 2024a. Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling. In *ACM SIGGRAPH 2024 Conference Papers (SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3641519.3657497

Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. 2024b. DragDiffusion: Harnessing Diffusion Models for Interactive Point-Based Image Editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 8839–8849. doi:10.1109/CVPR52733.2024.00844

Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. 2018a. Mastering Sketching: Adversarial Augmentation for Structured Prediction. *ACM Trans. Graph.* 37, 1 (Jan. 2018), 11:1–11:13. doi:10.1145/3132703

Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. 2018b. Real-Time Data-Driven Interactive Rough Sketch Inking. *ACM Transactions on Graphics* 37, 4 (Aug. 2018), 1–14. doi:10.1145/3197517.3201370

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. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=nJfylDvgzlq

Jaskirat Singh, Cameron Smith, Jose Echevarria, and Liang Zheng. 2022. Intelli-Paint: Towards Developing More Human-Intelligible Painting Agents. In *ECCV 2022* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 17 pages. doi:10.1007/978-3-031-19787-1_39

SmilingWolf. 2022. WD14 ViT Tagger. https://huggingface.co/SmilingWolf/wd-v1-4-vit-tagger-v2.

Yiren Song, Shijie Huang, Chen Yao, Hai Ci, Xiaojun Ye, Jiaming Liu, Yuxuan Zhang, and Mike Zheng Shou. 2024. ProcessPainter: Learning to Draw from Sequence Data. In *SIGGRAPH Asia 2024 Conference Papers (SA '24)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3680528.3687596

Jianchao Tan, Marek Dvorožňák, Daniel Sýkora, and Yotam Gingold. 2015. Decomposing Time-Lapse Paintings into Layers. *ACM Trans. Graph.* 34, 4 (July 2015), 61:1–61:10. doi:10.1145/2766960

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. *Advances in neural information processing systems* 29 (2016).

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-Guided Text-to-Image Diffusion Models. In *ACM SIGGRAPH 2023 Conference Proceedings (SIGGRAPH '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3588432.3591560

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-Video Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Zijian Zhang Rox Min Zuozhuo Dai Jin Zhou Jiangfeng Xiong Xin Li Bo Wu Jianwei Zhang Kathrina Wu Qin Lin Aladdin Wang Andong Wang Changlin Li Duojun Huang Fang Yang Hao Tan Hongmei Wang Jacob Song Jiawang Bai Jianbing Wu Jinbao Xue Joey Wang Junkun Yuan Kai Wang Mengyang Liu Pengyu Li Shuai Li Weiyan Wang Wenqing Yu Xinchi Deng Yang Li Yanxin Long Yi Chen Yutao Cui Yuanbo Peng Zhentao Yu Zhiyu He Zhiyong Xu Zixiang Zhou Zunnan Xu Yangyu Tao Qinglin Lu Songtao Liu Daquan Zhou Hongfa Wang Yong Yang Di Wang Yuhong Liu Weijie Kong, Qi Tian and along with Caesar Zhong Jie Jiang. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. https://arxiv.org/abs/2412.03603

Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2024. ToonCrafter: Generative Cartoon Interpolation. *ACM Trans. Graph.* 43, 6 (Nov. 2024), 245:1–245:11. doi:10.1145/3687761

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. 2024. Inversion-Free Image Editing with Natural Language. (2024).

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 18381–18391. doi:10.1109/CVPR52729.2023.01763

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* (2024).

Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. 2018. Two-Stage Sketch Colorization. *ACM Transactions on Graphics* 37, 6 (Dec. 2018), 1–14. doi:10.1145/3272127.3275090

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Paris, France, 3813–3824. doi:10.1109/ICCV51070.2023.00355

Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. 2017. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Transactions on Graphics (TOG)* 9, 4 (2017).

Amy Zhao, Guha Balakrishnan, Kathleen M Lewis, Frédo Durand, John V Guttag, and Adrian V Dalca. 2020. Painting many pasts: Synthesizing time lapse videos of paintings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8435–8445.

Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. 2021. Stylized Neural Painting. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15684–15693. doi:10.1109/CVPR46437.2021.01543