VINCIE: UNLOCKING IN-CONTEXT IMAGE EDITING FROM VIDEO

Anonymous authors
Paper under double-blind review

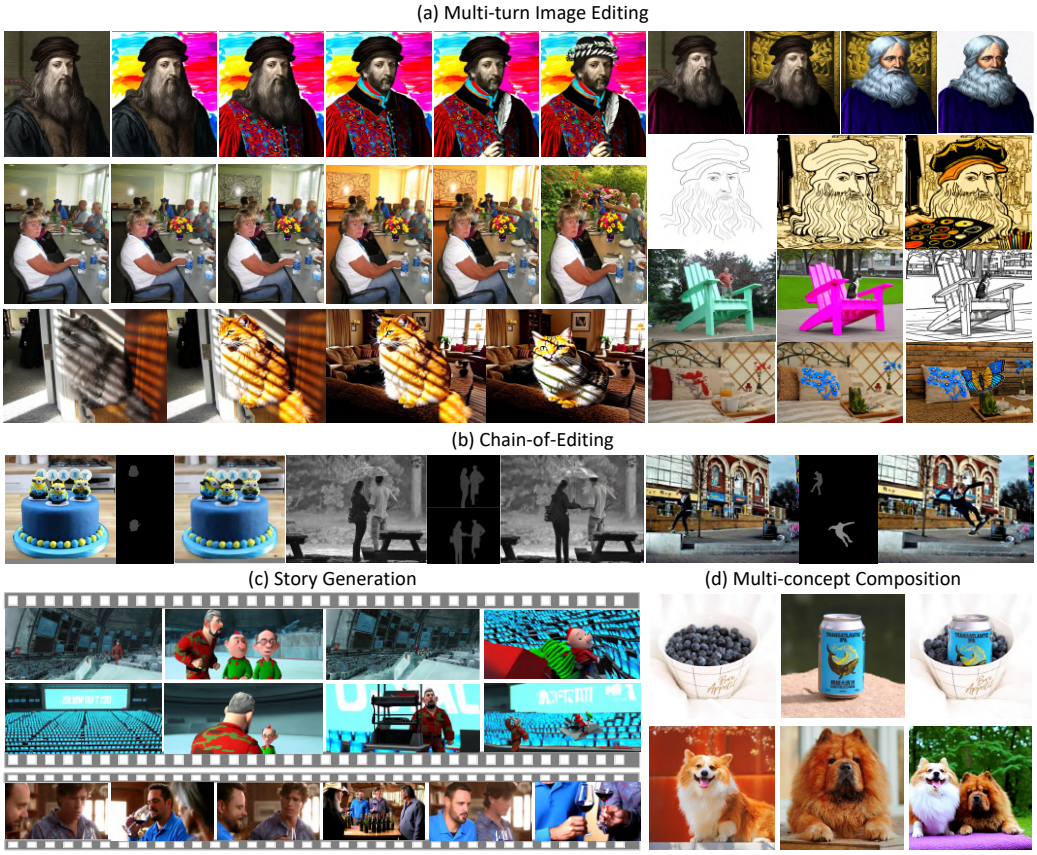

Figure 1: By learning from videos, our method could attain universal in-context editing and generation abilities to handle various practical creation scenarios.

ABSTRACT

In-context image editing aims to modify images based on a contextual sequence comprising texts and images. Existing methods typically depend on task-specific pipelines and expert models (*e.g.*, segmentation and inpainting) to curate training data. In this work, we explore whether an in-context image editing model can be learned directly from videos. Toward this end, we introduce a scalable approach to annotate videos as interleaved multimodal sequences. To effectively learn from this data, we design three proxy tasks: next-image prediction, current segmentation prediction, and next-segmentation prediction. Additionally, we propose a novel multi-turn image editing benchmark to advance research in this area. Extensive experiments demonstrate that our model exhibits strong in-context image editing capabilities and achieves state-of-the-art results on two multi-turn image editing benchmarks. Despite being trained exclusively on videos, our model also shows promising abilities in multi-concept composition, story generation, and chain-of-editing applications. Code is available at: https://anonymous.4open.science/r/VINCIE-11669/.

1 INTRODUCTION

Recent research has devoted significant effort to image editing, which enables users to generate images that closely follow editing instructions provided in text prompts. The performance of image editing models largely depends on the high-quality training data, typically composed of three elements: an input image, a text prompt describing the desired modification, and the corresponding edited image (Brooks et al., 2023; Shi et al., 2024a; Xiao et al., 2024; Wei et al., 2024; Han et al., 2024b; Xia et al., 2024; Liu et al., 2025). To collect such paired image data at scale, various methods have been proposed, including generating image grids (Wu et al., 2025c), leveraging diffusion denoising processes (Brooks et al., 2023), and developing specialized models or tools to extract before-and-after image pairs from the web (Hertz et al., 2022; Zhuang et al., 2024; Boesel & Rombach, 2024).

Very recently, the problem of *in-context image editing* (OpenAI, 2025b) has garnered growing interest in the research community. In this setting, a target image is generated based on a contextual sequence of text prompts and previously generated images. Unlike single-turn image editing, in-context image editing supports multi-turn interactions, enabling users to iteratively refine images while maintaining visual consistency throughout the editing process. A key challenge lies in acquiring contextualized training data that includes coherent sequences of text and images, Existing approaches to mine single-turn image editing (Brooks et al., 2023; Wu et al., 2025c; Hertz et al., 2022; Zhuang et al., 2024; Boesel & Rombach, 2024) struggle to construct meaningful long-form content that is capable of capturing the dependencies and evolving intent that emerge over multiple editing steps. The lack of contextualized, quality training data remains a significant barrier to progress in this area of research.

In this paper, we approach in-context image editing from a different perspective and investigate the following research question: *Can a meaningful in-context image editing model be learned solely from videos, without using any standalone images?* Our intuition is that videos, as a rich source of multimodal information, inherently contain a long duration of visual dynamics that might facilitate the learning of multi-turn interactions. For instance, changes within a scene, such as objects entering or exiting the frame, shifts in camera focus, or character actions, provide implicit cues for learning operations like addition, removal, and modification in image editing.

To this end, we propose an approach that natively learns transitions from video data, named Video-driven IN-Context Image Editing (**VINCIE**). Unlike conventional image editing methods that rely on separately collected pairs of pre- and post-editing images for training, we choose not to alter the video, *i.e.*, we train on native video data (only natural videos as the source of visual modality), but instead provide the model with detailed annotations that describe the transitions or actions occurring within the scene. Since our method eliminates the need for paired data collection and relies solely on video, it can be trivially scaled using the vast amount of video data readily available on the web.

Specifically, we first sample a few coherent frames from a video scene, annotate the visual transitions, and identify Regions of Interest for editing (RoEs) using a pretrained Vision-Language Model (VLM). Additionally, we employ Grounding-DINO (Liu et al., 2024b) and SAM2 (Ravi et al., 2024) to generate RoE segmentation masks based on textual descriptions of the transitions. This process establishes our training samples, which capture context and form an interleaved multimodal sequence. Next, we train a Diffusion Transformer (Peebles & Xie, 2023) with full attention as our primary implementation and additionally design a variant with block-wise causal attention, which applies bidirectional attention within each modality (frame, text, and segmentation mask) and causal attention across modalities. Both variants are compared to provide a direct assessment of their differences.

Finally, to enhance the model's learning of contextual dependencies, we design three proxy tasks: (1) next-image prediction, which serves as the primary task in training; (2) current segmentation prediction, which enables the model to understand which regions have changed; and (3) next segmentation prediction, which prepares the model to anticipate where changes are likely to occur.

Extensive experiments show that our model, trained solely on video data, demonstrates strong in-context image editing capabilities and outperforms existing baselines on the multi-turn image editing tasks. Scaling up the model and training data leads to substantial performance gains—for example, the success rate at the challenging 5-turn editing increases from 5% to 22% when scaling the training data from 0.25M to 10M sessions—demonstrating the scalability of our approach enabled by native video data. Notably, to the best of our knowledge, this is the first work to demonstrate the feasibility

of learning an in-context image editing model solely from video data, while also showcasing the scalability benefits of this approach.

We find that our model can learn disentangled representations of visual changes (*e.g.*, object appearance/disappearance, posture shifts, and orientation changes) purely from patterns inherent in video data. It also demonstrates reasonable generalization to scenarios that are less common in natural video, such as background changes, attribute modifications, and multi-concept compositions. As an additional benefit, our model can be used for generating consistent frames for storytelling through in-context editing.

2 Related Work

Data Construction Methods for Image Editing. Constructing image editing datasets requires first designing clear and diverse editing instructions that articulate the intended visual modifications. Based on these instructions, paired image examples are then created, consisting of original images and their corresponding edited versions that reflect the specified transformations. Single-turn image editing methods (Hertz et al., 2022; Brooks et al., 2023; Sheynin et al., 2024; Shi et al., 2024a; Zhao et al., 2024; Wei et al., 2024; Hui et al., 2024; Yang et al., 2024b; Jin et al., 2024) use pre-trained off-the-shelf models (Ramesh et al., 2022; Rombach et al., 2022; Brown et al., 2020; Sauer et al., 2024) to construct paired data for image editing. For example, InstructPix2Pix (Brooks et al., 2023) leverages GPT-3 (Brown et al., 2020) for generating editing instructions and Stable Diffusion v1.5 (Rombach et al., 2022) for paired image data generation. UltraEdit creates editing instructions using LLMs and combines grounding models (Kirillov et al., 2023; Liu et al., 2024b) with SDXL-Turbo (Sauer et al., 2024) to produce region-based editing samples. Our approach relies on learning transitions from videos without manually crafted paired data pipelines, bringing impressive scalability.

Learning from Video for Image Generation. Video Frames naturally exhibit consistency across characters, objects, and scenes, which has inspired recent efforts to construct source and target images from sampled video frames. Leveraging such frame-based data has proven beneficial for enhancing consistency in image generation tasks, such as instructive image editing (Chen et al., 2024d; Krojer et al., 2024), interactive image editing (Zhang et al., 2025a; Shi et al., 2024b), streamlining image editing (Alzayer et al., 2024), and object-level image customization (Chen et al., 2024c). The most recent work, *e.g.*, RealGeneral (Lin et al., 2025) and UES (Chen et al., 2024a), explored the temporal in-context consistency within video foundation models (Yang et al., 2024c) for universal image generation and editing. Despite notable progress, existing methods typically rely on only two frames per video, overlooking richer, long-range contextual information. Furthermore, they often depend on task-specific data construction pipelines (Chen et al., 2024d; Zhang et al., 2025a; Chen et al., 2024c), limiting their universality and scalability. In this work, we propose constructing session-wise data with long, interleaved image-text context from native videos, and leverage it for pre-training or mid-training to learn the inherent consistency and transformations in abundant multimodal sequences.

3 Methodology

3.1 Interleaved Multimodal Sequence Construction

Figure 2 shows an overview of our data construction pipeline. Starting with a video, we sparsely sample K frames $(I_0, \ldots, I_K)$ and use a vision-language model (VLM) to generate textual visual transitions T_i describing the change from frame I_i to I_{i+1}. To better capture the Regions-of-interest for editing (RoEs), we additionally annotate segmentation masks M_i and M_{i+1}, which identify the changing objects in I_i and I_{i+1}, respectively. Combining these elements, we construct the multimodal sequence $(I_0, T_0, T_{m0}, M_{00}, T_{m1}, M_{01}, I_1, \ldots, I_K)$. T_{m0} and T_{m1} are predefined prompts such as "generate the mask of changing areas in the source image" and "generate the mask of changing areas in the target image".

Frame Sampling. We use a hybrid sampling strategy: 1) *Equal-interval sampling*, which selects frames at fixed time intervals (*e.g.*, 3 sec), and 2) *Fixed-frame sampling*, which uniformly samples a fixed number (*e.g.*, $2 \leq n \leq 6$) of frames regardless of video duration. This approach is used to capture both subtle object-level changes and significant scene-level transitions.

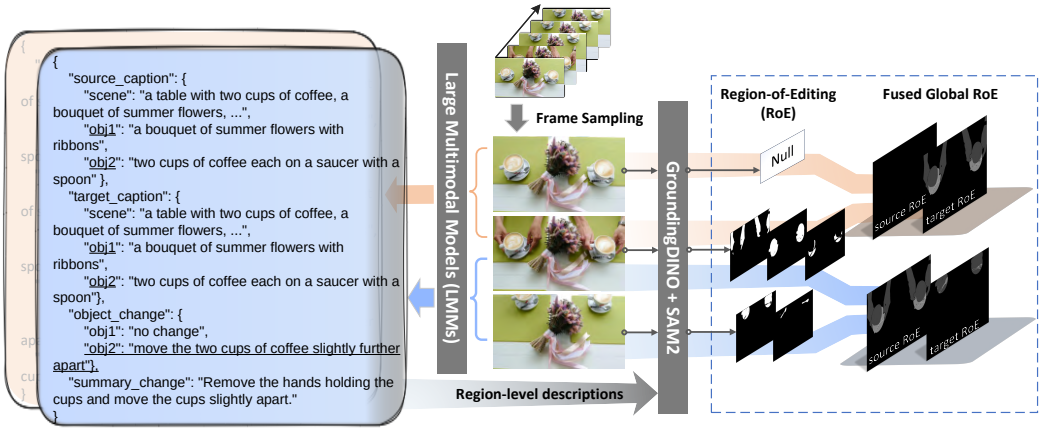

Figure 2: Our session data construction pipeline. We use a VLM to annotate the visual transitions. We then use the generated textual descriptions to prompt GroundingDINO+SAM2, extracting segmentation masks for the edited regions.

Visual Transition Annotation. To describe visual transitions between frames, we use chain-of-thought (CoT) prompting (Wei et al., 2022) to instruct a VLM to perform visual transition annotation: 1) generate detailed and coherent descriptions of each frame from multiple aspects (*e.g.*, characters, objects, attributes, interactions, scenes, and environments); 2) identify semantic and visual differences between the two frames from the above aspects; 3) and summarize all the differences into a concise, instruction-style statement T_i suitable for guiding editing. Unlike existing interleaved datasets (Zhu et al., 2023; Laurençon et al., 2023; Chen et al., 2024b) derived from web documents and retrieval tools, our dataset is built from native videos, ensuring stronger textual and visual coherence.

Segmentation Annotation and Encoding. We explicitly annotate Regions-of-Editing (RoEs) in both adjacent frames I_i and I_{i+1}. Specifically, we leverage region-level descriptions (*i.e.*, characters and objects) in the visual transition annotation as input to GroundingDINO (Liu et al., 2024b) and SAM 2 (Ravi et al., 2024) for extracting segmentation map's. Based on the region-level difference annotations, we determine which regions undergo visual transitions, *i.e.*, RoEs, and construct corresponding global maps by fusing local maps from the current and next session images.

3.2 Model Architecture

As illustrated in Fig. 3, our model is built upon a Diffusion Transformer (DiT) architecture, initialized from a video foundation model. We represent the interleaved input sequence as $S = (I_0, T_0, \ldots, T_{M-1}, I_M)$, where T_i denotes the textual editing instruction at turn-i, and I_i represents either an image or a segmentation mask.

As our focus is on the in-context image editing task, we optimize the model by maximizing the likelihood of the next image prediction:

$$\log p(S) = \sum_{i=1}^{M} \log p(I_i \mid I_0, \ldots, T_{i-1}, I_{i-1}) \tag{1}$$

where the conditional probability is modeled using flow-matching in the latent space, an objective commonly used in diffusion model for text-to-image (Rombach et al., 2022; Esser et al., 2024; Labs, 2024; Podell et al., 2023) and text-to-video (Singer et al., 2022; Wan et al., 2025; Hong et al., 2022; Seawead et al., 2025) generation tasks. Each text instruction (T_i) and image (I_i) is encoded into latent tokens using a text encoder (*e.g.*, T5) and an image encoder (*e.g.*, VAE), respectively. The details about the text encoder and VAE are provided in the supplementary material.

Learnable <TURN> Tokens. We separate the interleaved input sequence S by modality into two groups: $S = (I_0, T_0, \ldots, T_{M-1}, I_M) \rightarrow T = (T_0, T_1, ..., T_{M-1}); I = (I_0, ..., I_M)$. Their latent tokens are concatenated together. Since the number of text tokens at each turn may vary, we introduce M special learnable tokens $\texttt{<TURN>}_i, i = 1, ..., M$ to mark the turn boundary, where $\texttt{<TURN>}_i$ is inserted before the latent tokens of T_i.

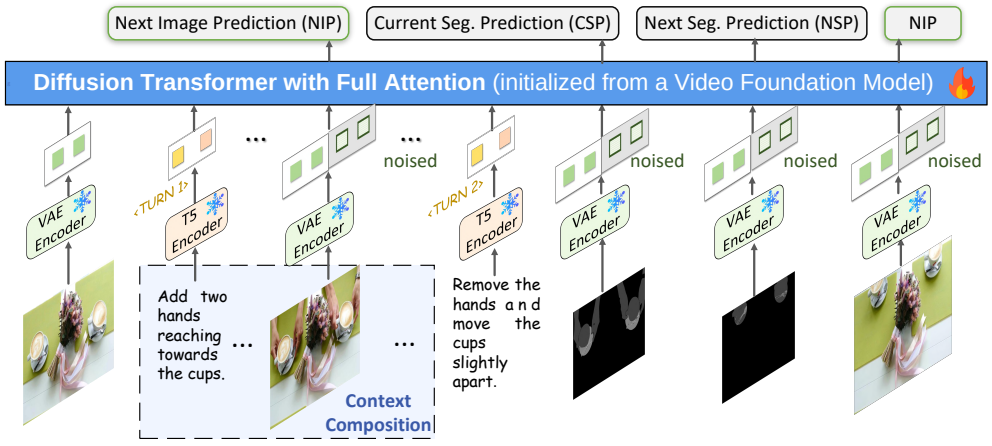

Figure 3: Model architecture. We apply a diffusion transformer framework (initialized from a video generative foundation model) with full attention to learn from the multimodal interleaved context, through three tasks (CSP, NSP, and NIP). Losses are only computed on noised tokens.

Separate Text and Image Position Embedding. We apply 1D RoPE (Su et al., 2024) to text tokens and 3D RoPE to image tokens. The starting positions are 0 for all dimensions. This separate RoPE design aligns with our pretrained MM-DiT model, where text and image tokens are positioned continuously. Position collisions are avoided as MM-DiT employs distinct weights for each modality, and the bias terms in the linear layers effectively act as modality-specific embeddings.

Attention. We employ two attention mechanisms in DiT and obtain two variants: (1) full attention over all tokens, as shown in Fig. 3, and (2) block-wise causal attention, where causality is enforced across blocks (*e.g.*, text or image) and bidirectional attention is applied within each block. Full attention enables comprehensive token interactions at a higher computational cost, while block-wise causal attention improves efficiency while maintaining causal structure. Additional details and discussions are provided in Appendix C.4.

Condition on Clean Context. We model the distribution of each image (except the first) using a diffusion loss, conditioned on an interleaved context. To enhance training efficiency, we concatenate the clean and noisy tokens of each image as model inputs, and apply an attention mask to ensure that each noisy image attends only to the clean representations of preceding images, as illustrated in Fig. 11.

3.3 Context Composition Learning

To facilitate effective ability transfer from segmentation modeling to image editing and generation, we unify image and segmentation modeling within a generative framework using the MSE-based diffusion loss in flow matching. Through interleaved context composition, our framework further unlocks multiple capabilities and supports a variety of corresponding tasks (see Fig. 12 for more details). Specifically, we augment Eqn. 1 by adding a random dropout operation Rd on the context, as shown in equation:

$$\log p(S) = \sum_{i=1}^{M} \log p(F_i \mid Rd(I_0, T_1), Rd(T_{m0}, M_{00}), Rd(T_{m1}, M_{01}), \dots) \tag{2}$$

where F_i can be either the target image, RoE mask[1] of the source image, or RoE mask of the target image. We ensure that the image or mask required to generate the target is always retained, while only the contextual images and texts are randomly dropped. The model is jointly learning three tasks:

- **Next Image Prediction (NIP)**. NIP is our primary in-context image editing task.

[1]In implementation, we treat segmentation masks as RGB images, by replicating the mask across all three channels, and then encode them using the VAE encoder to obtain the latents.

- **Current Segmentation Prediction (CSP)**. CSP enhances the model's *grounding* ability, enabling it to identify regions requiring edits while preserving consistency in other areas. This is particularly useful for local editing tasks such as removal, attribute changes, and replacements.
- **Next Segmentation Prediction (NSP)**. NSP improves the model's *controllable generation* by incorporating the next-frame segmentation map into the context, aiding in dynamic layout adjustments for scenarios like shape changes and movements.

By randomly combining different contexts and tasks, the model learns essential abilities such as grounding, controllable generation, and multi-concept composition, enabling versatile in-context image editing.

4 Experiments

4.1 Implementation Details

Data. Through the proposed scalable data construction pipeline, we collect and annotate about 10M session instances, with the number of images in each session from 2 to 20. For each session data, we consider RoE map with a probability of 80%. We apply a context drop rate with 20%, 70%, and 70%, to the current frame, current RoE map, and next RoE map, respectively. During inference, the sampling step is set to 50, the classifier-free guidance scale is set to 10. Using the proposed data construction pipeline, we collect and annotate about 10M session instances, each containing 2 to 20 images. During training, a RoE map is included with an 80% probability for each session. We apply context dropout rates of 20%, 70%, and 70% to the current frame, current RoE map, and next RoE map, respectively, with dropout applied independently at each turn. We use 50 sampling steps and set the classifier-free guidance scale to 10.

Model. We initialize our model with the weights of our in-house MM-DiT (3B and 7B), pre-trained on text-to-video tasks and architecturally similar to (Seawead et al., 2025; Kong et al., 2024). The 3B and 7B variants are optimized on session data for 15k and 40k steps, consuming approximately 30 and 150 hours on 256 H100 GPUs, respectively.

4.2 Multi-Turn Session Image Editing Benchmark

Existing benchmarks (Zhang et al., 2023a; Basu et al., 2023; Sheynin et al., 2024), such as MagicBrush (Zhang et al., 2023a), are constrained to basic editing operations, such as addition, replacement, removal, attribute modification, and background changes, and thus fall short of meeting practical user needs. Moreover, MagicBrush supports only up to three editing turns per session, with each turn treated in isolation, further diverging from real-world editing workflows. To address these limitations, we propose **MSE-Bench** (Multi-turn Session image Editing Benchmark), which comprises 100 test instances, each featuring a coherent five-turn editing session. MSE-Bench *expands the range of editing categories* to include more complex and realistic scenarios such as posture adjustment, object interaction, and camera view changes, as shown in Fig. 4. To better reflect user intent and practical applications, we also incorporate *aesthetic* considerations into the construction, encouraging progressive visual enhancement across turns.

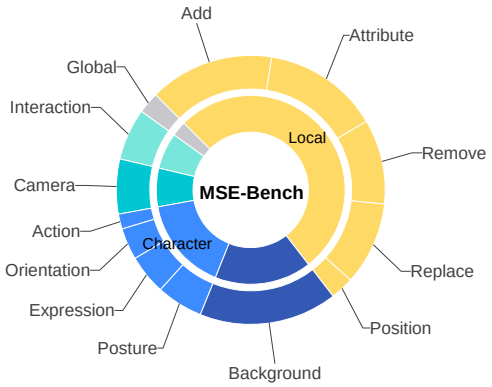

Figure 4: Category distribution of MSE-Bench. "others" includes expression, orientation, position, global, and action change.

For each editing instruction, multiple generated images may satisfy the user's request. Consequently, our benchmark does not provide ground-truth images. Instead, we use GPT-4o to evaluate whether the generated image successfully follows the instructions and remains consistent with the input image. The final score for each turn is computed by averaging the success rates across all samples.

Table 1: Performance comparison on MagicBrush (Zhang et al., 2023a) (multi-turn) for consistency (DINO and CLIP-I) and prompt following (CLIP-T). SFT means we carry out supervised fine-tuning. * denotes the use of context across all preceding turns. Entries by gray denote proprietary models.

Method	Turn-1			Turn-2			Trun-3		
	DINO	CLIP-I	CLIP-T	DINO	CLIP-I	CLIP-T	DINO	CLIP-I	CLIP-T
Instruct-Pix2Pix (Brooks et al., 2023)	0.514	0.727	0.270	0.397	0.674	0.268	0.335	0.646	0.263
MagicBrush (Zhang et al., 2023a)	0.826	0.901	0.278	0.756	0.863	0.277	0.718	0.834	0.271
HQEdit (Hui et al., 2024)	0.522	0.696	0.259	0.441	0.659	0.248	0.397	0.637	0.238
UltraEdit (Zhao et al., 2024)	0.755	0.852	0.289	0.706	0.827	0.278	0.683	0.810	0.266
ICEdit (Zhang et al., 2025b)	0.853	0.922	0.281	0.780	0.882	0.278	0.731	0.852	0.272
OmniGen (Zhang et al., 2025b)	0.874	0.924	0.273	0.718	0.851	0.264	0.586	0.786	0.261
OmniGen2 (Wu et al., 2025b)	0.863	0.919	0.285	0.777	0.869	0.280	0.716	0.832	0.278
Step1X-Edit (Liu et al., 2025)	0.852	0.915	0.288	0.785	0.875	0.286	0.743	0.840	0.277
Bagel (Deng et al., 2025)	0.845	0.912	0.286	0.767	0.873	0.292	0.723	0.844	0.286
Bagel* (Deng et al., 2025)	0.847	0.914	0.287	0.729	0.858	0.295	0.684	0.823	0.287
FLUX.1-Kontext (dev) (Batifol et al., 2025)	0.858	0.917	0.288	0.757	0.863	**0.296**	0.691	0.818	**0.291**
Qwen-Image-Edit (Wu et al., 2025a)	0.827	0.900	**0.292**	0.745	0.856	0.292	0.697	0.819	0.287
GPT Image 1* (OpenAI, 2025a)	0.805	0.875	0.293	0.708	0.820	0.300	0.666	0.789	0.292
Nano Banana* (DeepMind & Gemini, 2025)	0.886	0.933	0.287	0.811	0.896	0.294	0.773	0.867	0.291
Ours* (3B)	0.822	0.895	0.273	0.733	0.850	0.272	0.676	0.827	0.267
Ours* (3B) + SFT	0.852	0.917	0.283	0.739	0.861	0.291	0.667	0.814	0.290
Ours* (7B)	0.838	0.906	0.272	0.721	0.848	0.272	0.645	0.804	0.271
Ours* (7B) + SFT	**0.891**	**0.937**	0.283	**0.817**	**0.895**	0.289	**0.775**	**0.861**	0.286

4.3 Comparison with State-of-the-Arts

We evaluate our model on two multi-turn image editing benchmarks: MagicBrush (Zhang et al., 2023a) and our proposed MSE-Bench.

MagicBrush. Given its support for multi-turn editing, high-quality manual annotations, and close alignment with real-world editing needs, we first adopt MagicBrush to evaluate our method and compare against baselines. Tab. 1 reports quantitative results across three standard evaluation metrics: DINO, CLIP-I, and CLIP-T. First, our model, trained solely on interleaved video data, achieves performance comparable to SOTA methods UltraEdit and OmniGen, which rely on pairwise editing data, highlighting video data as a natural and effective source for image editing tasks. Second, with supervised fine-tuning on editing-oriented data, our method outperforms nearly all metrics, demonstrating that interleaved video data complements existing data creation approaches. Lastly, our model's advantages become increasingly evident with more edit turns, showcasing the benefits of learning from contextual video data.

MSE-Bench. Tab. 2 presents the multi-turn editing success rates as evaluated by GPT-4o. In this setup, the generated image at turn-i serves as the input for editing at turn-$i+1$. Consequently, failure at any turn propagates to subsequent turns. Existing academic methods perform poorly, with a success rate of **< 2%** at turn-5. In contrast, our method achieves a **25%** success rate at turn-5, demonstrating the advantages of our model and the use of native video data. However, our approach still falls short compared to proprietary models like GPT-4o, which benefit from significantly larger training datasets and model sizes. Even so, GPT-4o achieves only a **62.7%** success rate, highlighting the long-term value of our proposed benchmark for advancing multi-turn editing.

4.4 In-depth Analysis

In-Context Editing Mitigates Artifact Accumulation. Artifact accumulation, where artifacts become more pronounced with increasing editing turns, is a common issue in multi-turn editing (Sheynin et al., 2024). We observe this phenomenon as well (upper part of Fig. 6) when using our model as a single-turn editing method, *i.e.*, without incorporating context from previous turns. However, when all contexts are included as input, no artifacts are observed (lower part of Fig. 6).

Impact of Segmentation Prediction and Generation. As shown in Tab. 3, training with segmentation and generation as context enhances both consistency and multi-turn editing success rate. Notably, the substantial gain in consistency on MagicBrush (Zhang et al., 2023a) demonstrates the effectiveness of segmentation modeling, especially under the CoE strategy (CS $\rightarrow$ NS $\rightarrow$ I).

Impact of Context. Table 4 highlights the impact of context in multi-turn image editing.

Table 2: Performance comparison on MSE-Bench (editing success rate evaluated by GPT-4o). * denotes the use of context across all preceding turns. Entries by gray denote proprietary models.

Method	GPT-4o Evaluation				
	Turn-1	Turn-2	Turn-3	Turn-4	Turn-5
Instruct-Pix2Pix (Brooks et al., 2023)	0.520	0.130	0.110	0.083	0.060
MagicBrush (Zhang et al., 2023a)	0.707	0.300	0.213	0.170	0.087
HQEdit (Hui et al., 2024)	0.477	0.177	0.140	0.113	0.077
UltraEdit (Zhao et al., 2024)	0.673	0.230	0.173	0.113	0.067
ICEdit (Zhang et al., 2025b)	0.633	0.340	0.257	0.163	0.090
OmniGen (Xiao et al., 2024)	0.847	0.223	0.170	0.140	0.083
OmniGen* (Xiao et al., 2024)	0.853	0.188	0.160	0.125	0.065
OmniGen2 (Wu et al., 2025b)	0.847	0.393	0.327	0.263	0.133
Step1X-Edit (Liu et al., 2025)	0.937	0.540	0.420	0.300	0.140
Bagel (Deng et al., 2025)	0.967	0.650	0.613	0.550	0.413
Bagel* (Deng et al., 2025)	0.963	0.630	0.567	0.473	0.300
FLUX.1-Kontext (dev) (Batifol et al., 2025)	0.950	0.670	0.623	0.573	0.440
Qwen-Image-Edit (Wu et al., 2025a)	**0.980**	**0.737**	**0.667**	0.613	0.430
GPT Image 1 (OpenAI, 2025a)	0.963	0.690	0.673	0.637	0.557
GPT Image 1* (OpenAI, 2025a)	0.967	0.707	0.700	0.697	0.640
Nano Banana (DeepMind & Gemini, 2025)	0.987	0.773	0.753	0.727	0.627
Nano Banana* (DeepMind & Gemini, 2025)	0.997	0.773	0.757	0.730	0.643
Ours* (3B)	0.913	0.450	0.393	0.300	0.210
Ours* (3B) + SFT	0.913	0.533	0.497	0.443	0.330
Ours* (7B)	0.837	0.517	0.463	0.400	0.350
Ours* (7B) + SFT	0.950	0.693	**0.667**	**0.617**	**0.487**

Table 3: **Impact of segmentation (seg.) prediction and generation as context during training and inference on consistency** (CLIP-I and DINO on MagicBrush) **and success rate** (evaluated by GPT-4o). I: image generation. CS: current segmentation generation. NS: next segmentation generation. (This ablation study was conducted using an intermediate checkpoint, so the reported numbers may not be directly comparable to those in other tables.)

Train	Inference	MagicBrush (CLIP-I)			MagicBrush (DINO)			MSE-Bench (Success Rate by GPT-4o)				
		Turn-1	Turn-2	Turn-3	Turn-1	Turn-2	Turn-3	Turn-1	Turn-2	Turn-3	Turn-4	Turn-5
w/o Seg.	I	0.875	0.824	0.784	0.765	0.663	0.592	0.847	0.473	0.337	0.177	0.113
w/ Seg.	I	0.880	0.832	0.797	0.786	0.680	0.604	**0.887**	0.520	0.327	0.183	0.103
w/ Seg.	CS → I	0.886	0.832	0.801	0.797	0.687	0.622	0.873	**0.590**	**0.407**	**0.260**	**0.173**
w/ Seg.	NS → I	0.889	0.840	0.815	0.807	0.711	0.661	0.837	0.487	0.323	0.197	0.117
w/ Seg.	CS → NS → I	**0.890**	**0.847**	**0.823**	**0.814**	**0.724**	**0.679**	0.867	0.523	0.367	0.190	0.110

In Turn-1, where no prior context exists, adding a dummy context—comprising the original image and an instruction, In Turn-2 and Turn-3, where editing instructions and ground-truth images from previous turns are provided as context, adding a dummy context results in minimal improvements. "generate the same image," prepended before Turn-1—significantly improves performance. The L1 and L2 distances are nearly halved, indicating greater consistency between the generated image and the original image in unchanged areas, as these distances are measured pixel-wise. This is expected, as the existing context already provides sufficient information. These findings underscore the critical role of context in multi-turn image editing tasks.

Table 4: **Impact of context on multi-turn image editing with MagicBrush.** The "Dummy-Context" includes the original image and the instruction, "generate the same image." "History" refers to providing previous turns' ground-truth images as context. Results show that performance significantly improves when a reasonable context is included, emphasizing the importance of context in multi-turn image editing.

Method	L1↓	L2↓	DINO↑	CLIP-I↑	CLIP-T↑
Turn-1					
w/o Context	0.155	0.063	0.814	0.894	0.277
Dummy-Context	0.086	0.031	0.850	0.913	0.277
Turn-2					
w/o Context	0.159	0.067	0.834	0.902	0.279
History	0.099	0.038	0.845	0.909	0.278
Dummy-Context	0.087	0.033	0.869	0.922	0.280
Turn-3					
w/o Context	0.164	0.071	0.851	0.904	0.273
History	0.088	0.034	0.878	0.923	0.273
Dummy-Context	0.088	0.034	0.895	0.929	0.272

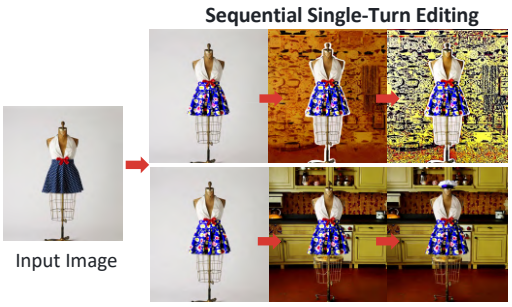

Figure 6: In-context editing mitigates artifact accumulation issue in sequential single-turn editing.

Figure 7: Subject position shift can be addressed by predicting segmentation mask first.

Scalability. Fig. 5 illustrates the editing success rate as a function of training data size. While the success rate at Turn-1 begins to saturate at 2.5M training samples, the success rate at later turns (*e.g.*, Turn-4 and Turn-5) exhibits a nearly log-linear increase with more training data. These results demonstrate the scalability of both our model and data construction pipeline.

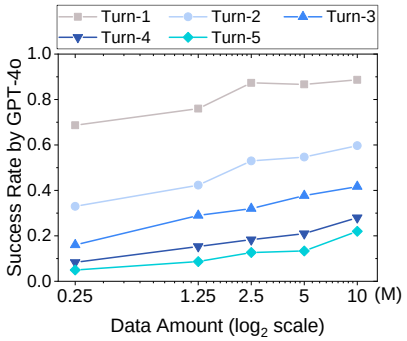

Figure 5: Editing success rates in 5 turns at various data scales.

Training on Native Video Data Introduces Addressable Subject Position-Shift. A key challenge when training on videos is the potential for subject position shifts across editing turns, as illustrated in the upper part of Fig.7. This issue arises from the natural movement of subjects over time in videos. However, incorporating segmentation prediction—where the model first predicts a mask before generating the target image—mitigates this drifting effect (see lower part of Fig.7). The segmentation mask enforces consistency in unedited regions, thereby reducing positional drift.

Effectiveness of Our Video Sequence Data. Table 5 demonstrates the impact of incorporating our video sequence data. Using the same pretrained model, training with our video sequence data increases success rates by **16.4%** and **21.0%** on Turn-1 and Turn-5, respectively, compared to training solely on specialized pairwise image editing data (Wei et al., 2024). The highest performance is achieved by first pretraining on our video sequence data, followed by supervised fine-tuning (SFT) on pairwise data, underscoring the effectiveness of our data for continual pretraining.

Table 5: Ablation study on MSE-Bench (GPT-4o evaluated success rate), to assess the impact of our video sequence data.

Training Data	Turn-1	Turn-2	Turn-3	Turn-4	Turn-5
pairwise	0.723	0.263	0.123	0.033	0.010
sequence	**0.887**	0.597	0.417	0.280	0.220
sequence → pairwise	0.880	**0.647**	**0.483**	**0.370**	**0.250**

4.5 Applications

Fig. 1 showcases several emerging capabilities that arise when training our model exclusively on video data. Notably, these abilities seem to develop implicitly, as they differ from the model's explicit training objectives:

- **Controllable Editing:** By including the segmentation mask of the region of interest in the context, users can achieve controllable editing by modifying the segmentation mask.
- **Multi-Concept Composition:** The model demonstrates the ability to compose multiple concepts together, even without explicit composition training data—a surprising emergent capability.
- **Story Generation:** Leveraging the consistent and extended context in video data, the model can generate coherent frames for storytelling through in-context editing.
- **Chain-of-Editing:** Each multi-turn editing session functions as a multimodal chain of thought, where the model interprets editing instructions, identifies regions of interest, generates RoI masks,

produces target images, and iterates the process. Our model reveals the potential of video data in modeling multimodal chains of thought.

5 Conclusion

In this work, we explore the research question: "Can an in-context image editing model be learned solely from videos?" To address this, we propose a learning framework that enables context-aware image generation directly from native videos. We introduce a scalable data construction pipeline that transforms videos into contextual multimodal sequences, comprising sparsely sampled frames, textual visual transition descriptions, and segmentation masks of regions of interest. To model this multimodal sequence, we train a DiT model using three proxy tasks: next-image prediction, current segmentation prediction, and next-segmentation prediction. Experimental results demonstrate that our model, trained exclusively on videos, exhibits strong in-context image editing capabilities and achieves state-of-the-art performance on multiple multi-turn image editing benchmarks. Additionally, our model showcases emerging abilities such as controllable editing, multi-concept composition, story generation, and multimodal chain-of-thought, highlighting the untapped potential of video data and the effectiveness of our proposed framework.

Ethics Statement

Our work on scalable, context-aware image editing has the potential to democratize creative tools, enhance accessibility, streamline media production, and advance intuitive human-AI collaboration. However, it also raises important concerns, including the risk of misuse for misinformation or manipulation, privacy issues from large-scale video data, potential biases in generated content, job displacement in creative industries, and increased environmental impact due to computational demands. Addressing these challenges will require careful dataset curation, privacy safeguards, bias mitigation, responsible deployment practices, and ongoing engagement with diverse stakeholders.

Reproducibility Statement

We have taken several steps to ensure the reproducibility of our work. An anonymous link (https://anonymous.4open.science/r/VINCIE-11669/) to the source code is provided, enabling replication of our implementation. The main text and appendix together provide comprehensive descriptions of the model design, training procedure, and evaluation protocol. Details on the dataset construction and preprocessing pipeline are presented in the appendix. These resources collectively ensure that readers can reproduce and validate our experimental results.

References

Johannes Ackermann and Minjun Li. High-resolution image editing via multi-stage blended diffusion. *arXiv preprint arXiv:2210.12965*, 2022.

Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. Magic fixup: Streamlining photo editing by watching dynamic videos. *ACM Transactions on Graphics*, 2024.

Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*, 2023.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Frederic Boesel and Robin Rombach. Improving image editing models with generative data refinement. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8861–8870, 2024.

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*, pp. 18392–18402, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Haodong Chen, Lan Wang, Harry Yang, and Ser-Nam Lim. Omnicreator: Self-supervised unified generation with universal editing. *arXiv preprint arXiv:2412.02114*, 2024a.

Songyan Chen and Jiancheng Huang. Fec: Three finetuning-free methods to enhance consistency for real image editing. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pp. 76–87. IEEE, 2023.

Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2406.10462*, 2024b.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

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*, pp. 6593–6602, 2024c.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. *arXiv preprint arXiv:2412.07774*, 2024d.

Gayoon Choi, Taejin Jeong, Sujung Hong, and Seong Jae Hwang. Dragtext: Rethinking text embedding in point-based image editing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 441–450. IEEE, 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Google DeepMind and Google Gemini. Nano banana (gemini 2.5 flash image) – image editing generation model, 2025. URL https://blog.google/products/gemini/updated-image-editing-model/. Official announcement: "Image editing in Gemini just got a major upgrade. Nano Banana is the latest upgrade to image generation in the Gemini app.".

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12736–12746, 2023.

Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7430–7440, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Peyman Gholami and Robert Xiao. Diffusion brush: A latent diffusion model-based editing tool for ai-generated images. *arXiv preprint arXiv:2306.00219*, 2023.

Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *CoRR*, 2023.

Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025.

Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6986–6996, 2024.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025.

Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4291–4301, 2024a.

Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024b.

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.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*, 2023.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

Ying Jin, Pengyang Ling, Xiaoyi Dong, Pan Zhang, Jiaqi Wang, and Dahua Lin. Reasonpix2pix: instruction reasoning dataset for advanced image editing. *arXiv preprint arXiv:2405.11190*, 2024.

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*, pp. 6007–6017, 2023.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2426–2435, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Chris Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulation. *Advances in Neural Information Processing Systems*, 37:38035–38078, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, et al. Zone: Zero-shot instruction-guided local editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6254–6263, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. *arXiv preprint arXiv:2503.10406*, 2025.

Yupei Lin, Sen Zhang, Xiaojun Yang, Xiao Wang, and Yukai Shi. Regeneration learning of diffusion models with rich prompts for zero-shot image translation. *arXiv preprint arXiv:2305.04651*, 2023.

Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6743–6752, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024b.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Jingyi Lu, Xinghui Li, and Kai Han. Regiondrag: Fast region-based image editing with diffusion models. In *European Conference on Computer Vision*, pp. 231–246. Springer, 2024.

Qi Mao, Lan Chen, Yuchao Gu, Zhen Fang, and Mike Zheng Shou. Mag-edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6842–6850, 2024.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G Derpanis, and Igor Gilitschenski. Watch your steps: Local image and scene editing by text instructions. In *European Conference on Computer Vision*, pp. 111–129. Springer, 2024.

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2063–2072. IEEE, 2025.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023.

Ivona Najdenkoska, Animesh Sinha, Abhimanyu Dubey, Dhruv Mahajan, Vignesh Ramanathan, and Filip Radenovic. Context diffusion: In-context aware image generation. In *European Conference on Computer Vision*, pp. 375–391. Springer, 2024.

Shen Nie, Hanzhong Allan Guo, Cheng Lu, Yuhao Zhou, Chenyu Zheng, and Chongxuan Li. The blessing of randomness: Sde beats ode in general diffusion-based image editing. *arXiv preprint arXiv:2311.01410*, 2023.

OpenAI. Introducing 4o image generation, 2025a. URL https://openai.com/index/introducing-4o-image-generation/.

OpenAI. Addendum to gpt-4o system card: Native image generation. *openai*, 2025b.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1104–1113, 2021.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 643–654, 2023.

Leigang Qu, Wenjie Wang, Yongqi Li, Hanwang Zhang, Liqiang Nie, and Tat-Seng Chua. Discriminative probing and tuning for text-to-image generation. *arXiv preprint arXiv:2403.04321*, 2024.

Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. Tiger: Unifying text-to-image generation and retrieval with large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. Silmm: Self-improving large multimodal models for compositional text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18497–18508, 2025b.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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*, pp. 10684–10695, 2022.

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.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.

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*, pp. 8871–8879, 2024.

Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing. *arXiv preprint arXiv:2411.06686*, 2024a.

Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8839–8849, 2024b.

Joonghyuk Shin, Daehyeon Choi, and Jaesik Park. Instantdrag: Improving interactivity in drag-based image editing. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–10, 2024.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18310–18319, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zeyi Sun, Ziyang Chu, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. X-prompt: Towards universal in-context image generation in auto-regressive vision language foundation models. *arXiv preprint arXiv:2412.01824*, 2024.

Linoy Tsaban and Apolinário Passos. Ledits: Real image editing with ddpm inversion and semantic guidance. *arXiv preprint arXiv:2307.00522*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*, 2023a.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18359–18369, 2023b.

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7378–7387, 2023.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025c.

Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing. *arXiv preprint arXiv:2412.17098*, 2024.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22428–22437, 2023.

Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *Advances in Neural Information Processing Systems*, 36:10359–10384, 2023.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18381–18391, 2023a.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*, 2024a.

Ling Yang, Bohan Zeng, Jiaming Liu, Hong Li, Minghao Xu, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing. *arXiv preprint arXiv:2405.14785*, 2024b.

Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3190–3199, 2023b.

Zhen Yang, Ganggui Ding, Wen Wang, Hao Chen, Bohan Zhuang, and Chunhua Shen. Object-aware inversion and reassembly for image editing. *arXiv preprint arXiv:2310.12149*, 2023c.

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*, 2024c.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

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*, 36:31428–31449, 2023a.

Shiwen Zhang, Shuai Xiao, and Weilin Huang. Forgedit: Text guided image editing via learning and forgetting. *arXiv preprint arXiv:2309.10556*, 2023b.

Yabo Zhang, Xinpeng Zhou, Yihan Zeng, Hang Xu, Hui Li, and Wangmeng Zuo. Framepainter: Endowing interactive image editing with video diffusion priors. *arXiv preprint arXiv:2501.08225*, 2025a.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025b.

Zhongping Zhang, Jian Zheng, Zhiyuan Fang, and Bryan A Plummer. Text-to-image editing by image information removal. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5232–5241, 2024.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36: 8958–8974, 2023.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2024.

Siyu Zou, Jiji Tang, Yiyi Zhou, Jing He, Chaoyi Zhao, Rongsheng Zhang, Zhipeng Hu, and Xiaoshuai Sun. Towards efficient diffusion-based image editing with instant attention masks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7864–7872, 2024.

CONTENTS

A The Use of Large Language Models

We acknowledge that large language models (LLMs) were employed to assist in the preparation of this manuscript. Their use was restricted to grammar checking, language refinement, and enhancing clarity and fluency of the text. In addition, LLMs were applied in a limited capacity to support minor debugging and syntactic corrections of code snippets.

B Additional Related Work

Image Editing. Building on advances in foundational image generation models (Huang et al., 2025; Ramesh et al., 2022; Saharia et al., 2022; Esser et al., 2024), image editing has achieved remarkable progress. Techniques now enable a wide range of edits, including zero-shot editing (Li et al., 2024; Huang et al., 2023; Wu & De la Torre, 2023; Han et al., 2024a; Chen & Huang, 2023), changing object classes (Kim et al., 2022; Xu et al., 2023; Ackermann & Li, 2022; Yang et al., 2023c; Tsaban & Passos, 2023; Gholami & Xiao, 2023; Brack et al., 2024; Nie et al., 2023) and faces (Ding et al., 2023), free-form text-based modifications (Brooks et al., 2023; Hertz et al., 2022; Lin et al., 2023; Dong et al., 2023; Zhang et al., 2023b; Kawar et al., 2023; Guo & Lin, 2024; Zhang et al., 2024; Sheynin et al., 2024; Wei et al., 2024; Shi et al., 2024a; Wang et al., 2023a; Li et al., 2023; Mirzaei et al., 2024; Miyake et al., 2025), mask-based edits (Wang et al., 2023b; Xie et al., 2023; Couairon et al., 2022; Zou et al., 2024; Mao et al., 2024), point dragging (Mou et al., 2023; Shin et al., 2024; Liu et al., 2024a; Lu et al., 2024; Choi et al., 2025), and reference image-guided transformations (Song et al., 2023; Goel et al., 2023; Yang et al., 2023a). A series of recent works (Yang et al., 2023b; Wu et al., 2023; Xiao et al., 2024; Najdenkoska et al., 2024; Sun et al., 2024) enables edits conditioned on multiple text and images. Our work focuses on in-context image editing (OpenAI, 2025b), where edits are conditioned on a contextual sequence of text and *previously generated* images. Moreover, we explore learning from native video data, unlike existing methods that use hand-crafted synthesized data.

C Implementation Details

C.1 Data Details

The training videos are sourced from a wide spectrum of domains, including stock footage, films, documentaries, etc. We split the raw videos into both single-shot clips and multi-shot scene videos. We also pre-process the raw videos by using different filtering strategies to keep high-quality videos, including logo detection, black border detection, and aesthetic estimation.

As described in Sec.3.1, we adopt two frame sampling strategies: equal-interval sampling and fixed-frame sampling. As illustrated in Fig.8, these approaches jointly ensure both the diversity and temporal stability of visual dynamics—two key factors for effective training of in-context image editing models.

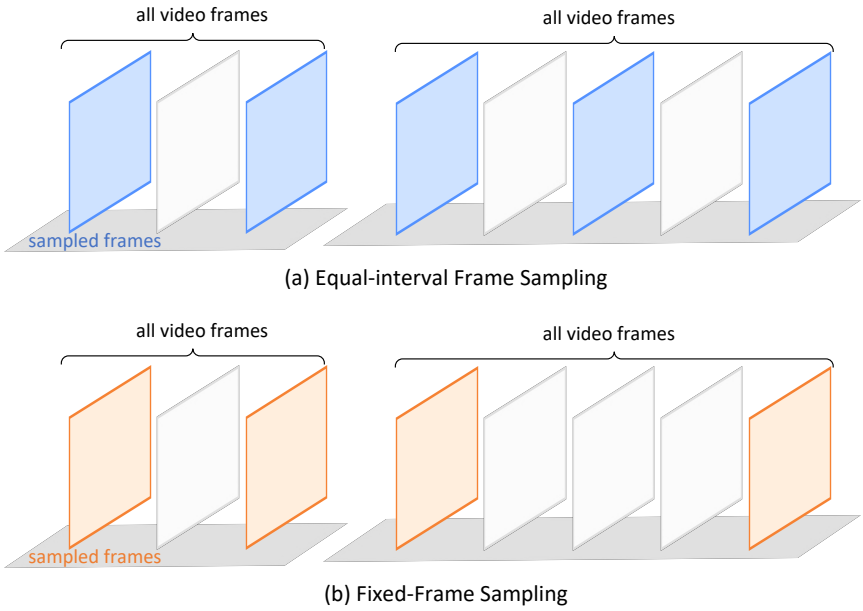

Figure 8: Two ways of frame sampling: (a) equal-interval sampling and (b) fixed-frame sampling.

C.2 Visual Transition Annotation

Instruction for Visual Transition Annotation

Imagine that you are an image editing assistant who wants to edit the first image to the second image. I will provide you two frames from a video clip as the source and target images. The caption of the raw video clip is: {}

Your task is to summarize how you intend to achieve this image editing task by providing detailed but brief text instructions, from the following guidelines:
1. Understand the two images first, and describe the two frames in detail and coherently. Please include the details of the environment, main subjects, their appearances, and main features.
2. Describe the main characters and objects and their appearances. Do not mention the real name entities. Follow the format such as: {"char1": "a woman with blonde hair wearing a red jacket", "char2": "a girl wearing a floral dress", "obj1": "a green apple", ...}
3. Highlight the semantic and visual differences between the two images in detail.
4. Provide only factual descriptive differences based on observable content. Avoid words or phrases that suggest speculation or assumptions, such as "likely", "possibly", or "appear to".
5. Avoid elliptical referential pronouns, such as "the same, frame 1, frame 2, the first image, the second image, ... ".

An editing instruction should include:
1. main character change, including appearance, disappearance, position, action, expression, pose, orientation, ... (*e.g.*, "make the person smile")
2. object change, including appearance, disappearance, position, count, relationship, layout, ... (*e.g.*, "add a dog beside the person")
3. attribute change, including color, texture, material, shape, size, depth, dynamics, ... (*e.g.*, "make the person's hair red")
4. interaction change, including the interaction between characters, objects, and the environment. (*e.g.*, "make the person hold the dog")
5. global change, including background, atmosphere, environment, style, weather, season, lighting, ... (*e.g.*, "make the weather dark")
6. camera change, including orbiting, dolly-in, dolly-out, pan-left, pan-right, tilt-up, tilt-down.
7. others

Output Format: You should output a json file to include the following information:
Frame1 Caption: <describe the first image/frame, characters and objects in detail>
Frame2 Caption: <describe the second image/frame, characters and objects in detail>
Character Change: <the detailed character and attribute change>
Object Change: <the detailed object and attribute change>
Global Change: <the detailed global change>
Camera Change: <the detailed camera change>
Other Change: <the detailed other change>
Summary Change: <a comprehensive but brief user editing instruction to achieve the editing>
Your output should be a JSON file in one row (without any format), which looks like:
{"frame1_caption": {"scene": str, "char1": str, "char2": str, ..., "obj1": str, "obj2": str, ...}, "frame2_caption": {"scene": str, "char1": str, "char2": str, ..., "obj1": str, "obj2": str, ...}, "character_change": {"char1": str, "char2": str, ...}, "object_change": {"obj1": str, "obj2": str, ...}, "global_change": str, "camera_change": str, "other_change": str, "summary_change": str}

To bridge the semantic gap between two sampled frames, we use our in-house LMM to annotate visual transitions, as introduced in Sec.3.1. The instruction used during annotation is shown above, and Fig. 10 presents example annotations to illustrate their quality.

C.3 Segmentation Mask Annotation and RoE Construction

The proposed visual transition annotation framework leverages an LMM to generate multi-level annotations, ranging from local concepts to global scene descriptions. As illustrated in Fig.2, we first use character and object descriptions from the source and target frames as query inputs to GroundingDINO(Liu et al., 2024b) to obtain object detection results. These detections are then passed to SAM 2 (Ravi et al., 2024) to extract segmentation masks for the corresponding local concepts. Guided by the annotated local changes, we identify and fuse the objects or characters undergoing transitions to construct the final RoEs.

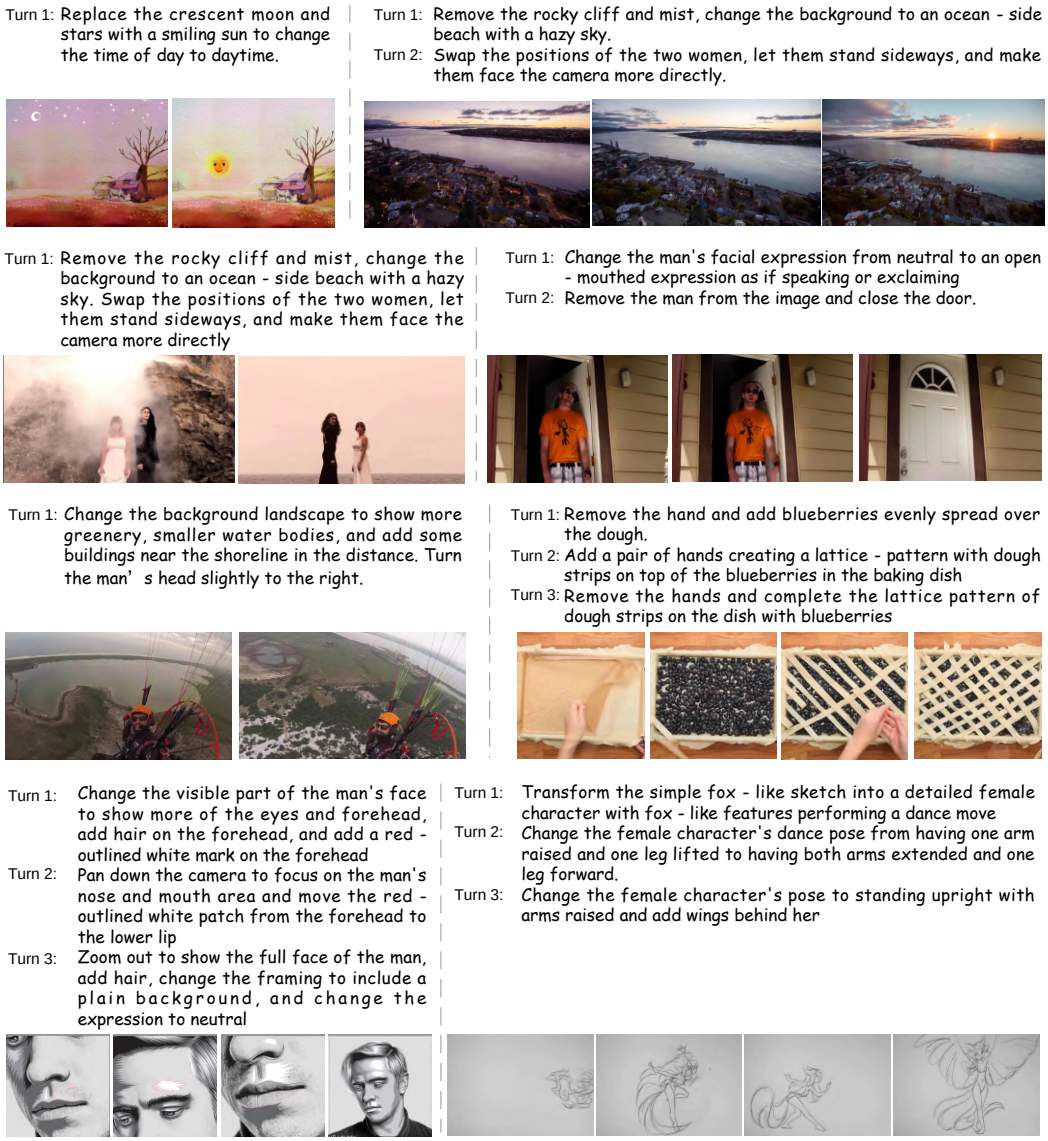

Figure 9: Examples (1/2) of visual transition annotation performed by our in-house large multimodal model.

Turn 1: Stop the man's hand - gesturing and close his mouth 1,
Turn 2: Add a curved stick to the man's left hand and make him gesture with his right hand.
Turn 3: Change the man's hand gesture from a general gesture to a rock - on hand gesture with the arm raised higher

Turn 1: Change the boy's action to running with one arm extended towards the basketball and move the basketball to in front of the boy on the ground, and change the boy's orientation to face more towards the left.
Turn 2: Change the boy's action to standing upright and looking forward, move the basketball to the boy's right hand, and add a sun on the right side of the sky.
Turn 3: Add a black X on the boy's shorts and change his pose to holding the basketball up to his face

Turn 1: Remove the two women and add a white armchair with a blanket, a small black round table, a floor lamp with a white shade, and a potted plant.
Turn 2: Add a woman with red hair, wearing a yellow short - sleeved shirt and black pants, standing and facing away from the camera with her right hand raised slightly to the room scene.
change the woman's action from walking and gesturing to standing and touching the patterned curtain with both hands
Turn 3: Add a woman with blonde hair sitting on the armchair, holding a white cup and raising her hand.
Turn 4: Change the first woman's action to standing and holding a white cup and smiling. Add two white cups, one in each woman's hand.

Turn 1: Move the man closer to the SUV such that he is opening the rear door with his right hand, and change the SUV's rear door to be open.
Turn 2: Edit the image to transition the man's position from standing outside the rear door of the SUV to being partially inside the vehicle, bent over.
Turn 3: Remove the man getting into the SUV and close the rear door.
Turn 4: Move the white SUV further down the path and close the rear right door.

Figure 10: Examples (2/2) of visual transition annotation performed by our in-house large multimodal model.

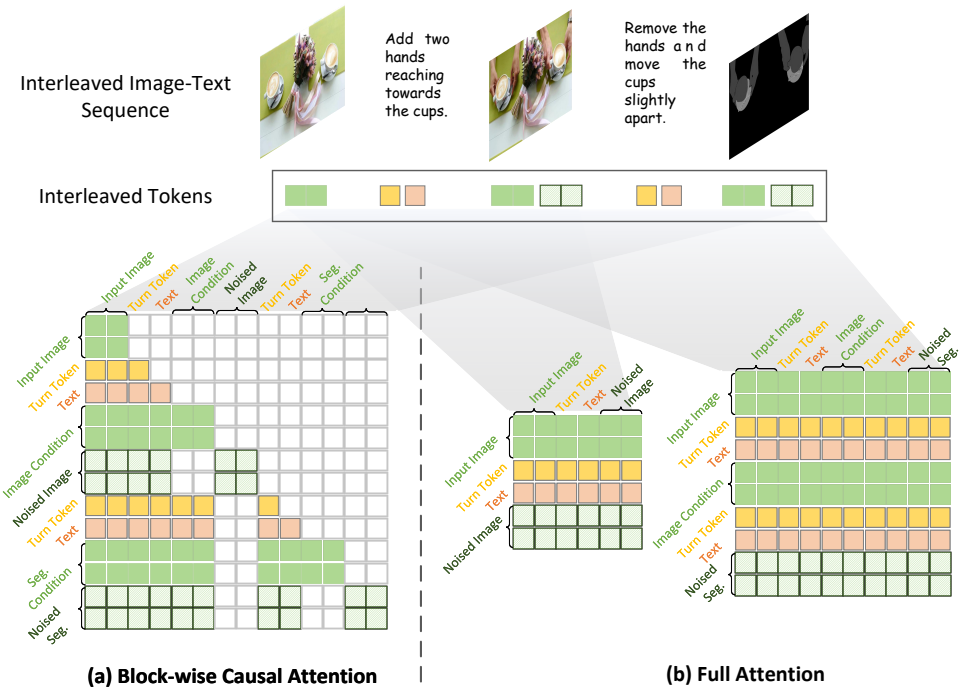

Figure 11: Implementation of (a) block-wise causal attention and (b) full attention.

C.4 Model Architecture

Variational Autoencoder. Following prior work (Yu et al., 2023), we adopt the encoder in a pretrained VAE to embed each image into the latent space separately for efficient computation. Specifically, it compress raw pixels with shape $(H, W, 3)$ into a (h, w, c)-shape latent representation, with downsampling ratios as $d_h = \frac{H}{h}$ and $d_w = \frac{W}{w}$ for height and width, respectively, and the latent channel c. The decoder in VAE aims to transform latent representations generated by the DiT back into the pixel space during inference.

Text Encoder. We employ the pretrained Flan-T5 as the text encoder to separately encode the prompt in each turn, and then concatenate all the embedding with inserting turn embeddings in between. Specifically, to make the model better discriminate different turns, we define a special turn token $\texttt{<TURN>}_i$ for the i-th turn, and introduce a learnable turn embedding for each one, which is inserted before the prompt embedding in the i-th turn.

Full Attention and Block-wise Causal Attention. We show the comparison between full attention and block-wise causal attention, and the condition strategy of clean context in block-wise causal attention, in Fig. 11.

C.5 Composition of Input Conditions and Output

In Fig. 12, we enumerate all seven context compositions supported by our method, detailing the interleaved input conditions, the corresponding outputs, the learning objectives, and the specific capabilities unlocked by each composition.

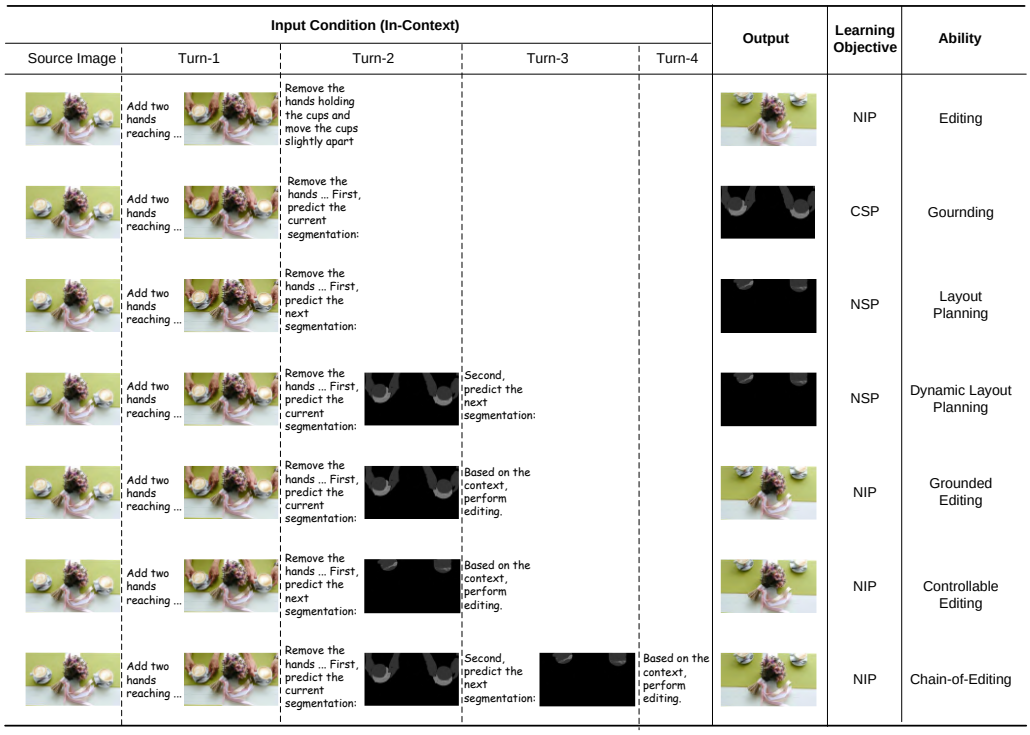

Figure 12: Context composition supported by our method.

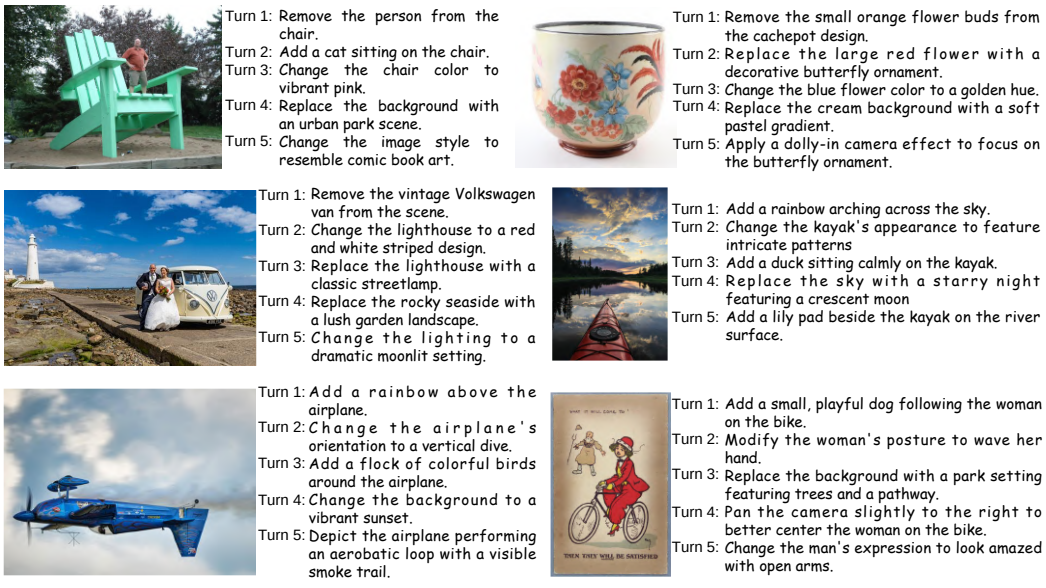

Figure 13: Multi-turn image editing examples of MSE-Bench.

C.6 Details of MSE-Bench

Instruction for Evaluation of Multi-turn (turn-i) Image Editing on MSE-Bench

Assume you are an expert in evaluating multi-turn image editing. In this task, a user interacts with an image editing system across multiple turns. At the first turn, the user provides a source image and an editing prompt. The system returns the edited image. In each subsequent turn, the user supplies a new prompt, and the system generates a new image based on the output from the previous turn. Your goal is to evaluate how successfully the editing instruction of the LAST turn has been executed.
You will be given user editing prompts and images: the first image is the original source image, and the next are the edited results from each turn for each prompt. You should focus more on the last prompt and the last edited image, but you may also consider the previous prompts and images as context.
The user editing prompts are: {}

Please follow these evaluation rules:
1. Last-turn Evaluation: For the last turn, you should first assess the result based on two criteria by giving a reason: 1) prompt_following, does the last edited image fulfill the last user's editing prompt? 2) consistency: Are the untouched parts of the last result image consistent with the input reference (the source image at the first turn, or the result image at the previous turn)?
2. Scoring: Based on the reason, you assign scores for "prompt_following" and "consistency".
From scale 0 to 10:
A "prompt_following" score from 0 to 10 will be given based on the editing success of prompt following. (0 indicates that the scene in the last edited image does not follow the last editing instruction at all. 10 indicates that the scene in the last edited image follow the last editing instruction perfectly.)
A "consistency" score from 0 to 10 will rate the degree of overediting in the last edited image. (0 indicates that the scene in the last edited image is completely different from the original. 10 indicates that the last edited image can be recognized as a minimally edited yet effective version of the original.)
3. Return your results in a JSON structure, following this format:
{{"reason": "...", "prompt_following": int, "consistency": int}}

The source images for our constructed multi-turn image editing benchmark, MSE-Bench, are sampled from MS-COCO (Lin et al., 2014) and LAION-Aesthetics (Schuhmann et al., 2022). Specifically, we randomly sample 6,000 images from each dataset and employ GPT-4o to perform prompt imagination, guided by criteria such as editing reasonability, aesthetics, consistency, and coherence. To facilitate this, we define a set of editing operations (e.g., add, remove, replace) and design a series of rules to instruct GPT-4o to simulate realistic and coherent multi-turn editing prompts from real users' perspectives. The instruction used in this process is illustrated above. Following prompt generation, we conduct careful human filtering to remove low-quality cases, resulting in a final set of 100 high-quality, category-balanced examples that constitute MSE-Bench. Additional examples are shown in Fig.13.

C.7 Supervised Fine-Tuning

After training on the constructed interleaved data from native videos, *i.e.*, VINCIE-10M, we carry out supervised fine-tuning to align the model with downstream editing tasks. Specifically, all of our SFT data comes from open-sourced datasets, including:

- OmniEdit-Filtered-1.2M[2] (Wei et al., 2024),
- Web-Image-3 and GRIT-Entity-New splits from X2I-subject-driven[3] proposed in OmniGen (Xiao et al., 2024),
- X2I2-video-editing, X2I2-inpaint-editing, X2I2-in-context-generation, and X2I2-in-context-editing splits from X2I2[4] proposed in OmniGen (Wu et al., 2025b),
- SEED-Data-Edit-Part3[5] proposed by SEED-X (Ge et al., 2024).

[2] https://huggingface.co/datasets/TIGER-Lab/OmniEdit-Filtered-1.2M
[3] https://huggingface.co/datasets/yzwang/X2I-subject-driven
[4] https://huggingface.co/datasets/OmniGen2/X2I2
[5] https://huggingface.co/datasets/AILab-CVC/SEED-Data-Edit-Part2-32

Table 6: Human evaluation on MSE-Bench based on editing success rate. * indicates use of context. Entries by gray denote proprietary models.

Method	Human Evaluation				
	Turn-1	Turn-2	Turn-3	Turn-4	Turn-5
HQEdit (Hui et al., 2024)	0.170	0.073	0.020	0.003	0.000
UltraEdit (Zhao et al., 2024)	0.310	0.062	0.015	0.002	0.000
OmniGen (Xiao et al., 2024)	0.333	0.035	0.002	0.000	0.000
GPT-4o*	0.872	0.783	0.755	0.642	0.491
Ours*	0.661	0.500	0.323	0.209	0.070

D Additional Experimental Results

D.1 Human Evaluation on Multi-turn Image Editing

To further verify the effectiveness and superiority of the proposed method for multi-turn image editing, we conduct human evaluations to assess editing success rates. The results are reported in Tab. 6. These findings validate the benefits of training on native video data, combined with supervised fine-tuning on pairwise editing examples, in enhancing multi-turn editing performance.

D.2 Correlation Between GPT-4o and Human Evaluation

Table 7: Correlation between automatic metrics and human evaluation

Metric	**GPT-4o vs Human**	**CLIP-T vs Human**	**CLIP-I vs Human**
Pearson r	**0.4858** ($p = 0.0000$)	0.0817 ($p = 0.4191$)	-0.0549 ($p = 0.5875$)
Spearman ρ	**0.4644** ($p = 0.0000$)	0.0692 ($p = 0.4941$)	-0.0217 ($p = 0.8303$)
Kendall τ	**0.4154** ($p = 0.0000$)	0.0502 ($p = 0.4963$)	-0.0195 ($p = 0.7921$)

In our experiments (Sec.4), we primarily report GPT-4o evaluated success rates to assess multi-turn image editing performance. To validate the reliability of GPT-4o-based evaluation, we compute the correlation between GPT-4o scores and human judgments. As shown in Tab.7, we also compare other metrics such as CLIP-T and CLIP-I. The results demonstrate that GPT-4o correlates well with human evaluation, supporting its use as a reliable proxy for scoring multi-turn image editing.

D.3 Human Evaluation for VLM annotation

To further verify the validity of the proposed automatic interleaved data construction pipeline, we conducted a human evaluation for VLM annotation on accuracy and recall, as shown in Tab. 8. While current VLMs are imperfect, they offer a scalable data annotation solution, achieving a favorable balance between quality and scalability. Similar to prior works (Brooks et al., 2023), our work aims to explore continual pre-training on the constructed large-scale interleaved corpus, where data scale is critical and minor annotation noise is tolerable. Besides, we believe the proposed automatic data construction pipeline will become increasingly effective, with the rapid advancement of large multimodal models.

Table 8: Human evaluation for VLM annotation on 500 data instances randomly sampled from our training dataset.

Metric	Score
Accuracy	75.14%
Recall	69.06%

D.4 Additional Ablation Study

Impact of RoPE and Attention. Based on a video foundation model, VINCIE continues pre-training on the constructed interleaved text-image data. In the foundation

Table 10: Ablation study on MSE-Bench (GPT-4o evaluated success rate), to assess the impact of RoPE and Attention.

RoPE	Attention	Turn-1	Turn-2	Turn-3	Turn-4	Turn-5
text-then-image	full	**0.968**	**0.360**	**0.320**	0.238	0.160
interleaved	full	0.933	0.338	0.308	**0.245**	**0.183**
interleaved	block-causal	0.880	0.290	0.230	0.200	0.120

Table 9: Impact of segmentation (seg.) prediction and camera prompt engineering (PE), *i.e.*, inserting "[###CAMERA: None###]" before each user prompt, on consistency evaluated by CLIP-I and DINO scores on Magicbrush (Zhang et al., 2023a).

Train	Inference	CLIP-I Score			DINO Score		
		Turn-1	Turn-2	Turn-3	Turn-1	Turn-2	Turn-3
w/o Seg.	-	0.875	0.824	0.784	0.765	0.663	0.592
w/ Seg.	-	0.880	**0.832**	0.797	0.786	0.680	0.604
w/ Seg.	Camera PE	**0.884**	**0.832**	**0.798**	**0.798**	**0.681**	**0.612**

Table 11: Our method vs. In-context LoRA (Huang et al., 2024) with human evaluation on the benchmark introduced in LCT (Guo et al., 2025) for story keyframe generation. Evaluation is carried out from two aspects: prompt following and consistency.

Metric	Win	Fail	Tie
Prompt Following	55.10%	16.30%	28.60%
Consistency	43.80%	2.10%	54.20%

model, we first concatenate text tokens and video tokens into a sequence, perform Rotary Position Embedding (RoPE) (Su et al., 2024) on it, and then feed it to multiple MM-DiT (Esser et al., 2024) layers for video modeling. Full bidirectional attention is adopted in each layer for thorough intra-modal and cross-modal interaction. To explore the impact of RoPE and Attention, we design three variants and conduct a performance comparison on MSE-Bench, as shown in Tab. 10. Considering the foundation model has carried out large-scale pre-training on text-video data, it has attained strong prior knowledge based on text-then-image RoPE and full attention. This strategy achieves the best performance on Turn-1 to Turn-3. However, the interleaved RoPE gradually outperforms it as the sequence length increases. One reason is that the interleaved RoPE arranges the text and image more naturally than the trivial text-then-image strategy. Finally, block-causal attention performs the worst, which may be attributed to the limited modality interaction. However, block-causal attention shows strong potential, offering flexibility in next-block modeling, support for prefill decoding to enable efficient inference, and compatibility with LLMs, which we leave for future work.

Impact of Camera Motion in Training Video Data. In most editing scenarios, consistency is highly required. To delve into possible entanglement issues of camera and object movement, we have adopted a disentanglement learning strategy consisting of: 1) Explicit annotation of camera change (see the instruction in Sec. C.2); 2) Incorporation of camera prompt wrapped in special tokens, such as "[###CAMERA: pan-left###]", during training; 3) Use of static camera prompt, *i.e.*, "[###CAMERA: None###]", during inference. This strategy enables the model to disentangle camera movement from object dynamics, allowing flexible camera control based on application needs. The results shown in Tab. 9 verify its effectiveness in improving consistency.

D.5 Additional Performance Comparison on Story Keyframe Generation

Tab. 11 provides a quantitative comparison of story keyframe generation performance between our method and the recent method, *i.e.*, In-context LoRA Huang et al. (2024) on the benchmark introduced in LCT (Guo et al., 2025), serving as empirical evidence of the effectiveness of our approach. In this work, we aim to introduce a general video-driven learning framework to unlock in-context image editing and generation, with story keyframe generation being one potential application. Due to the limited time, we focus on multi-turn image editing, while more comprehensive evaluation of other capabilities, including story generation, is left for future work.

Figure 14: **Zero-shot** qualitative results of single-turn image editing on cases uncommonly present in video data. The model was only trained with interleaved session data from video, T2I data, and T2V data.

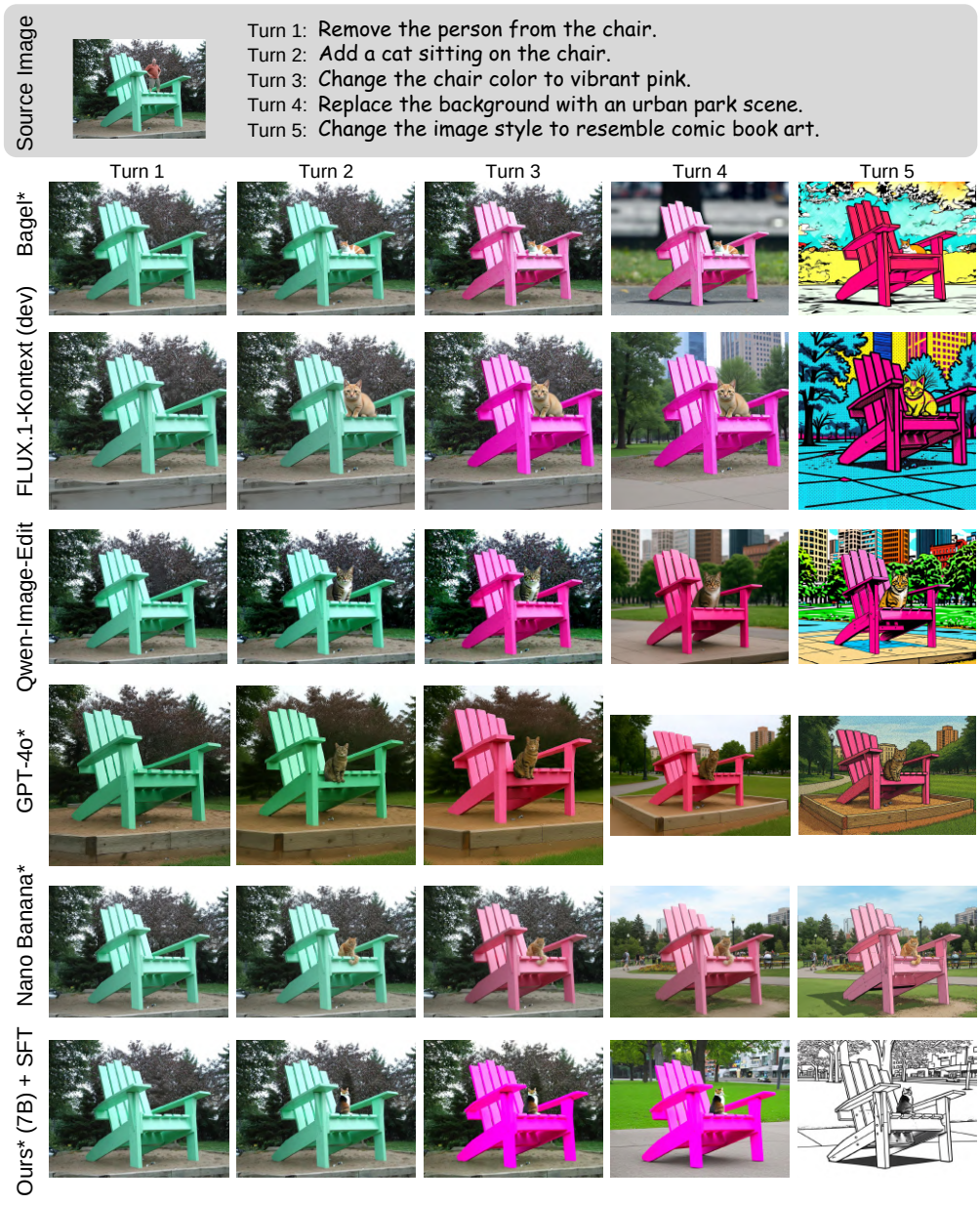

Figure 15: Qualitative comparison (1/4) between our method and recent baselines on MSE-Bench.

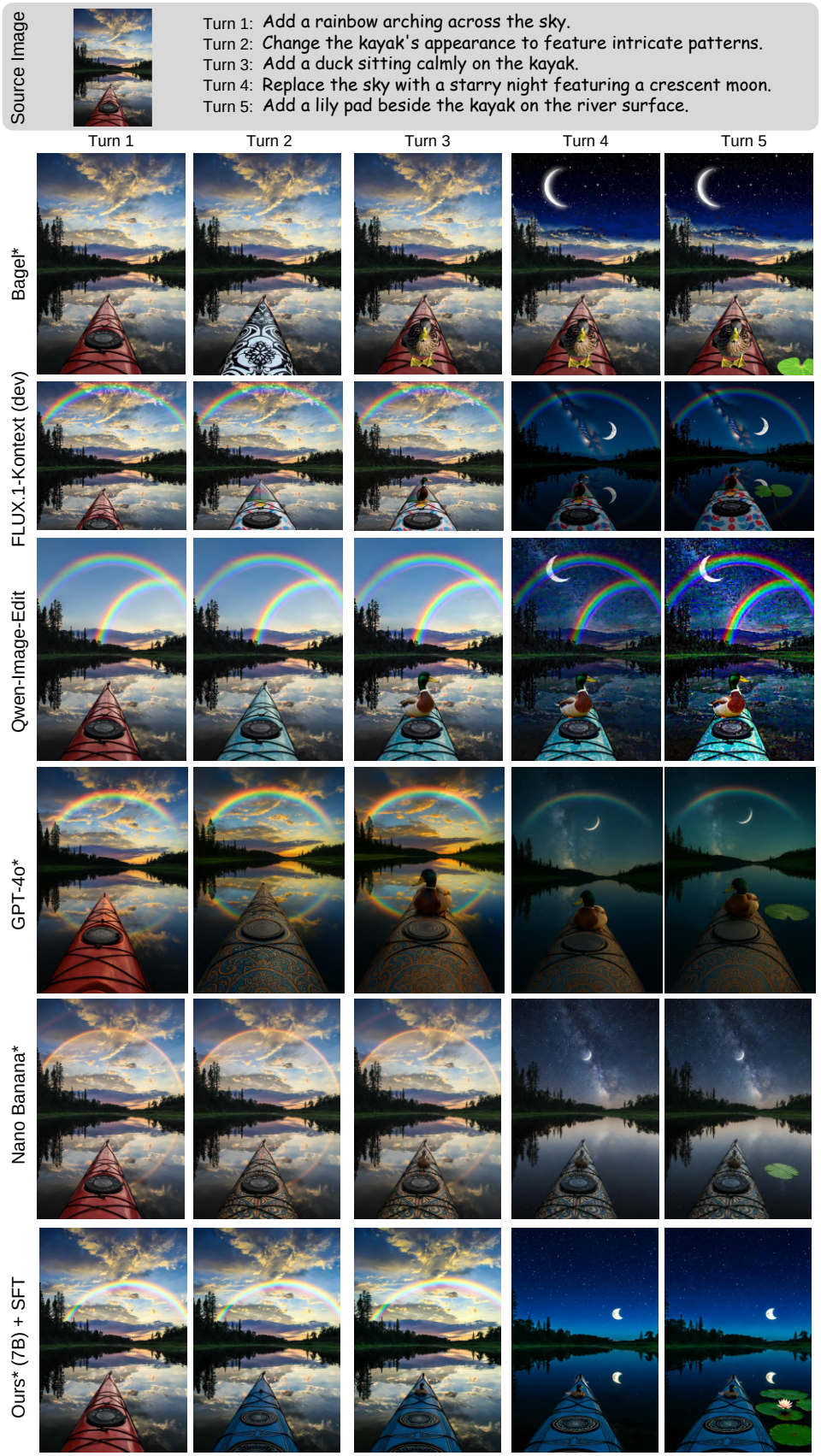

Figure 16: Qualitative comparison (2/4) between our method and recent baselines on MSE-Bench.

Figure 17: Qualitative comparison (3/4) between our method and recent baselines on MSE-Bench.

Figure 18: Qualitative comparison (4/4) between our method and recent baselines on MSE-Bench.

E Additional Application Examples

E.1 Image Editing

Single-turn Image Editing. In addition to common scene changes present in video data, we observe that our model generalizes well to uncommon cases, such as abrupt environmental shifts, complex style transfers, and material transformations (Fig. 14). This capability may arise for two reasons. First, although infrequent, such patterns (*e.g.*, environmental changes) are still present in our training corpus. Second, the model is initialized from a video foundation model that has been extensively pre-trained on both T2I and T2V data, enabling it to internalize high-level concepts such as style and material. These derived capabilities can be naturally transferred to the image editing setting.

Multi-turn Image Editing. As shown in Fig. 15, we compare our method with several baselines, including HQ-Edit (Hui et al., 2024), UltraEdit (Zhao et al., 2024), OmniGen (Xiao et al., 2024), and GPT-4o. The results reveal several key observations: 1) Most existing models suffer from error accumulation, leading to increasingly severe artifacts across editing turns. 2) These accumulated errors often degrade prompt-following performance, where the model fails to execute edits as instructed once artifacts dominate. 3) While GPT-4o—a strong proprietary model—achieves competitive results, it may exhibit inconsistencies in some cases compared to our method. 4) Overall, these comparisons highlight the effectiveness of training on native video data for achieving coherent and prompt-aligned multi-turn image editing. Additional qualitative examples are provided in Fig. 26, Fig. 27, Fig. 28, and Fig. 29, further demonstrating the strong prompt-following and consistency of our approach across multiple editing turns.

E.2 Multi-concept Composition

In-context Image Generation. In Fig. 19, we present qualitative results on in-context image generation for multi-concept composition, which requires both composition and strong identity preservation. These examples demonstrate that only training on video data (without any fine-tuning) can effectively unlock compositional capabilities, despite the rarity of such patterns in typical video content. This emergent behavior highlights the potential of video-based pre-training. Further scaling of model capacity, compute resources, and video data may enable the emergence of even more advanced capabilities.

In-context Image Editing. In addition, we conducted further supervised fine-tuning on the X2I2 (Wu et al., 2025b) dataset to explore more advanced multi-concept composition abilities. The qualitative results (Fig. 20) on in-context image editing indicate that even lightweight SFT substantially incentivizes more powerful compositional editing abilities, highlighting the effectiveness of our video-driven pre-training. Notably, compared with Fig. 19, **our fine-tuned model generalizes beyond object-centric concepts, such as background, color, and expression**, despite these concepts being uncommon in the fine-tuning dataset (Wu et al., 2025b).

E.3 Story Generation

Since our method is trained on native video data, it inherently captures the underlying storylines present in the sequences. As illustrated in Fig.21, we formulate story generation as a multi-turn image editing task, guided by transition prompts between key frames during inference. These examples showcase the model's ability to follow prompts while maintaining coherence and consistency across turns. When combined with existing long video generation methods(Guo et al., 2025), our approach has the potential to enhance top-down planning for generating coherent long-form story videos.

E.4 Chain-of-Editing

In Tab. 3, we show the effectiveness of chain-of-editing, *i.e.*, predicting segmentation maps before performing image editing. The predicted segmentation maps could be viewed as a kind of "thoughts". In Fig. 22, we show more qualitative results for challenging cases to demonstrate the effectiveness of CoE.

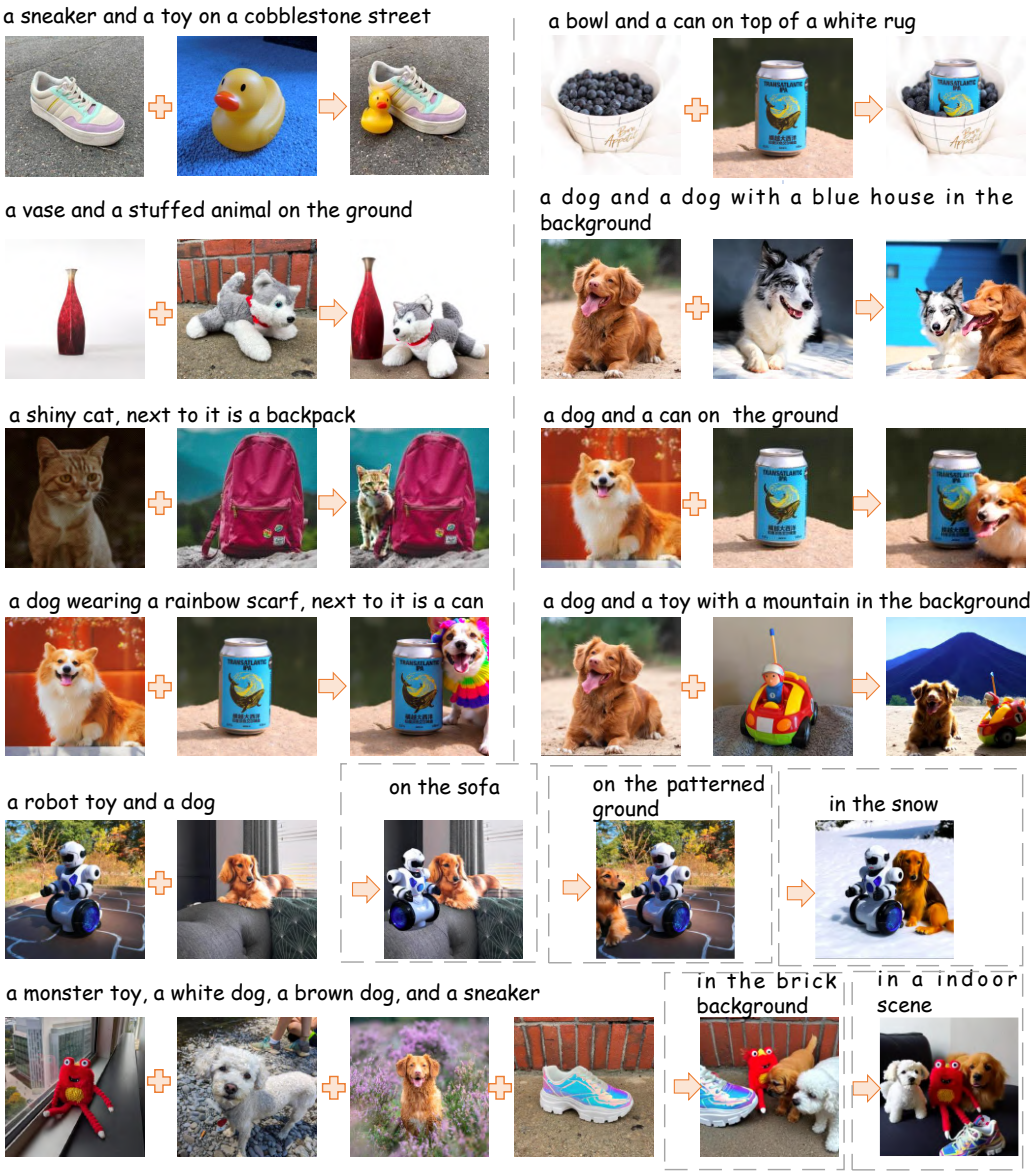

Figure 19: **Zero-shot** qualitative results of multi-concept composition (in-context generation) achieved by our method (without any fine-tuning).

Figure 20: Qualitative results of multi-concept composition (in-context editing) achieved by our method. Our model is fine-tuned on the X2I2 (Wu et al., 2025b) dataset, but the shown transferred concepts (*e.g.*, background, color, and expression) are uncommon in X2I2. It demonstrates the strong generalization ability conferred by pre-training on video-based interleaved data.

Character 1: A woman with long brown hair, wearing a light-colored top.

Turn 1: Transform the scene from a large robot crashing amidst debris with people scattering to a focused battle scene between two robots in a city street.
Turn 2: Transition from a dynamic battle scene involving robots in a damaged urban setting to a close-up of a woman operating a vehicle's controls, emphasizing personal struggle over large-scale action.
Turn 3: Make the woman's expression more desperate. Add a thin metallic object to her hand to indicate urgency in her actions.
Turn 4: Transition from an interior vehicle scene with Character 1 trying to start the car, to an outdoor chaotic city scene featuring a giant robot, soldiers, and a passing black car.
Turn 5: Shift from a smoky, chaotic city street scene with a towering robot and watching soldiers to a close-up, detailed view of the robot's face with glowing red eyes, eliminating the context and ambient elements.
Turn 6: widen the focus from the robot's face to include a scene of combat with intense explosions and debris, adding context and action to the static appearance of the robot in frame 1.
Turn 7: Decrease the flames and smoke to expose the robots and enhance their details in the frame.

Character 1: A man with curly hair wearing a patterned blue shirt.
Character 2: A man with short, slightly curly hair, a mustache, and a goatee, wearing a blue shirt.
Character 3: A person with long gray hair and a cowboy hat.

Turn 1: Shift focus from Character 1 to Character 2 holding the wine glass to his nose, with Character 1 slightly repositioned to the background.
Turn 2: Add Character 1 to the second frame, adjusting Character 2's pose to hold the wine glass similar to the second frame, while maintaining the indoor wine room's ambiance and lighting.
Turn 3: Widen the shot to include a third character with long gray hair and a cowboy hat behind the counter, and add several wine bottles on the counter to transform the scene from a close-up to a wide shot, encompassing a group tasting session.
Turn 4: Change from a wide shot in a wine tasting room to a medium close-up shot focusing on Character 1 and Character 2 with a blurred background, removing Character 3 and the visible wine bottles.
Turn 5: Shift the focus from the characters' upper body in Frame 1 to their hands swirling wine glasses in Frame 2, emphasizing the interaction with the wine.

Figure 21: More qualitative results of story generation achieved by our method.

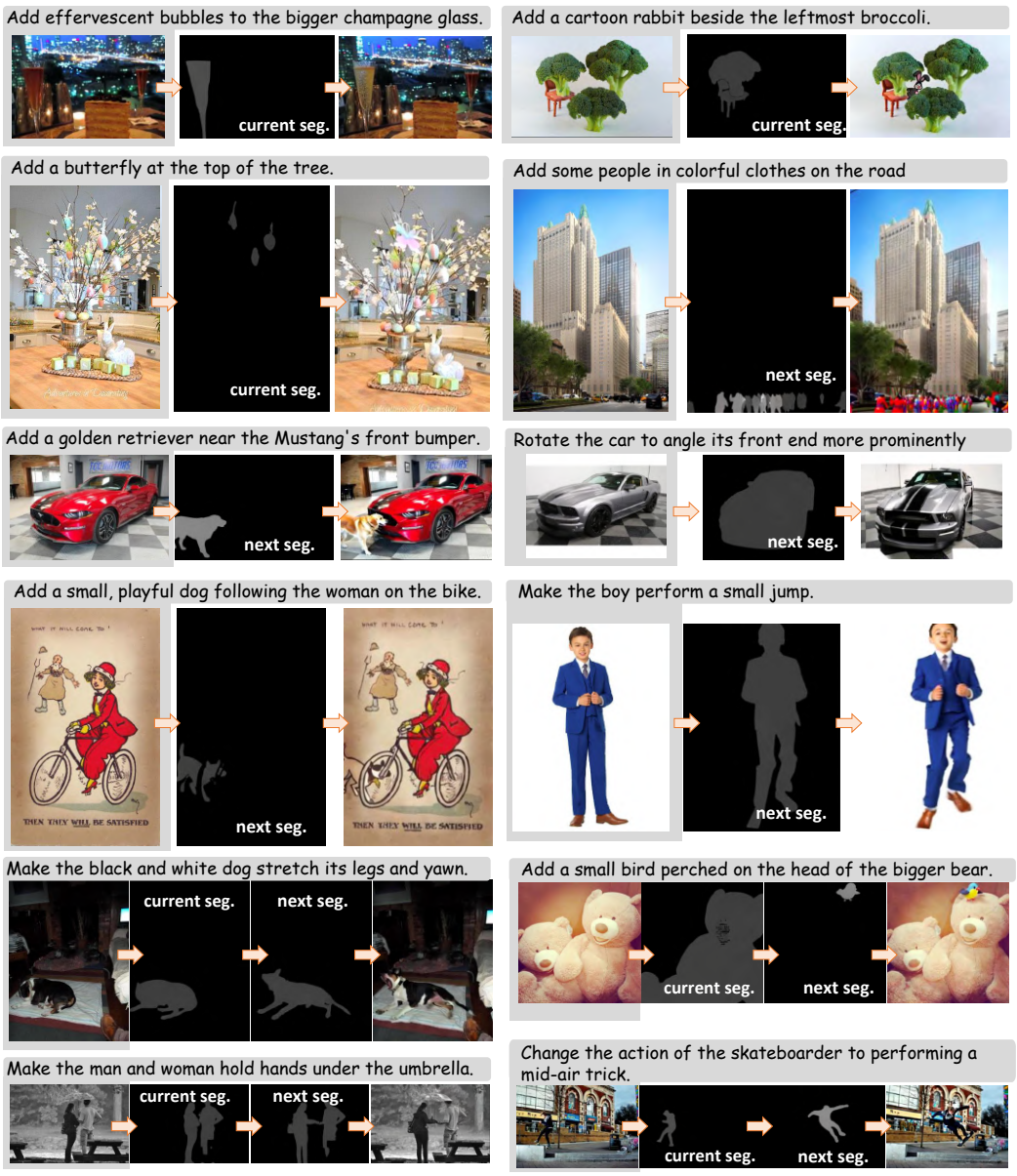

Figure 22: More qualitative results of Chain-of-Editing.

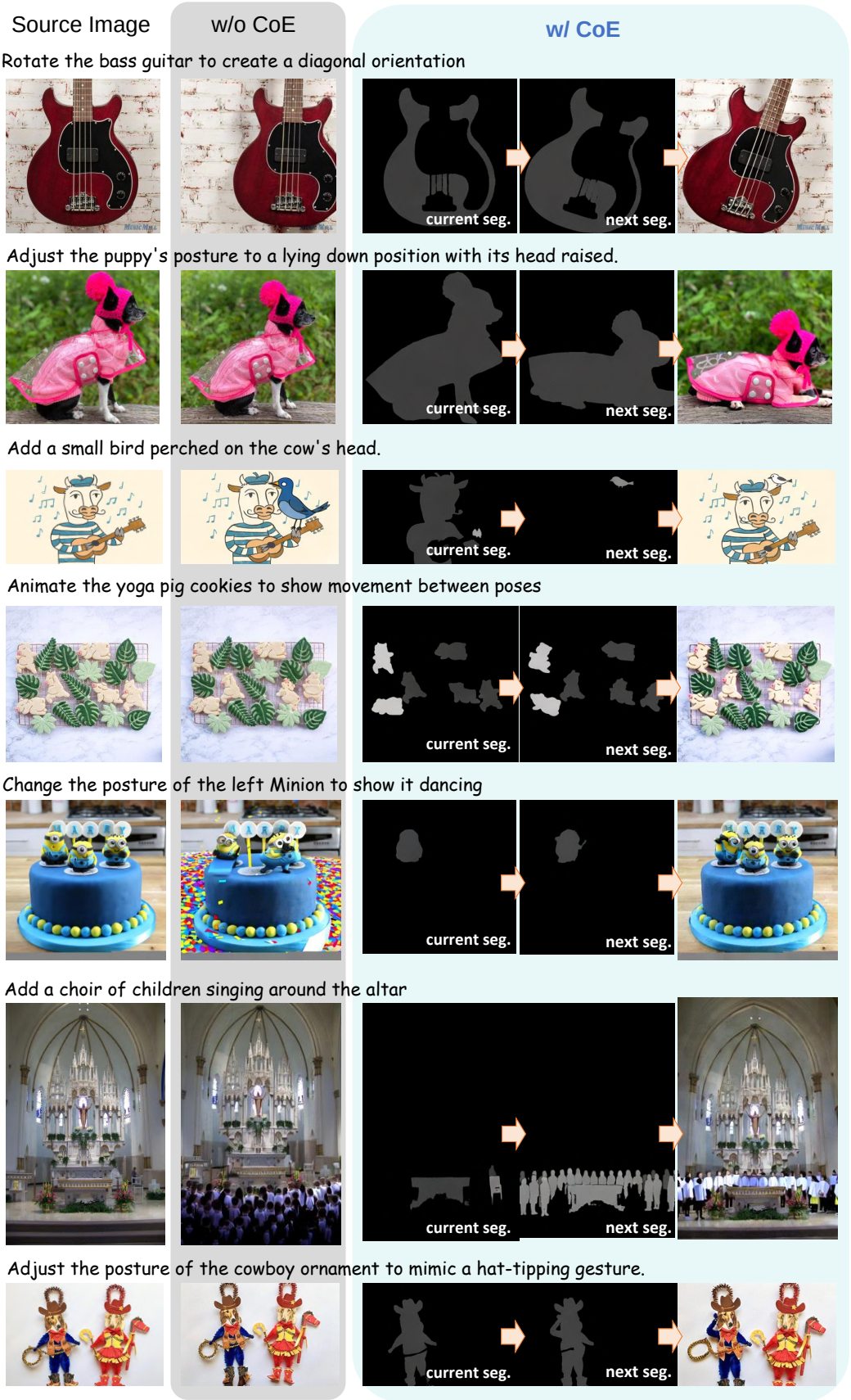

Figure 23: Qualitative comparison between w/o Chain-of-Editing (CoE) and w/ CoE.

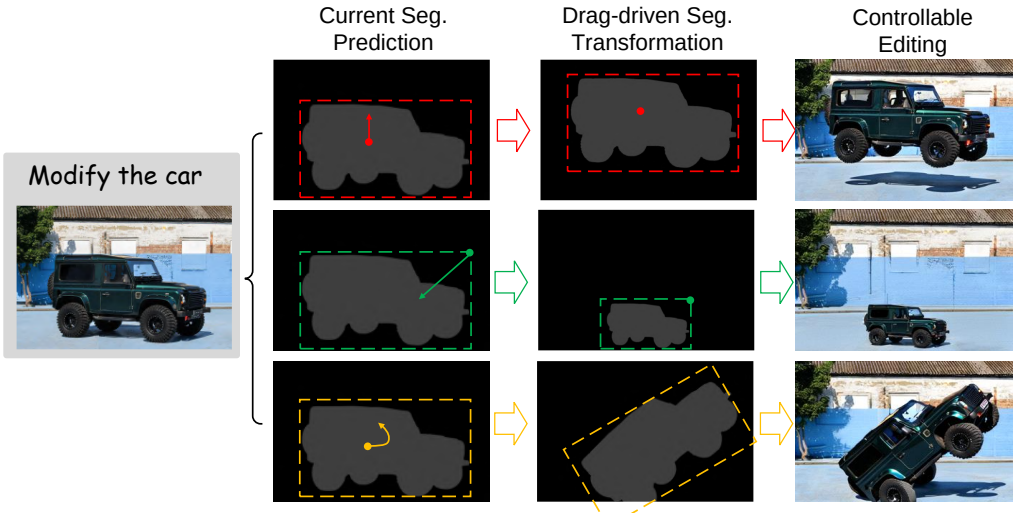

Figure 24: Qualitative results of drag-based image editing.

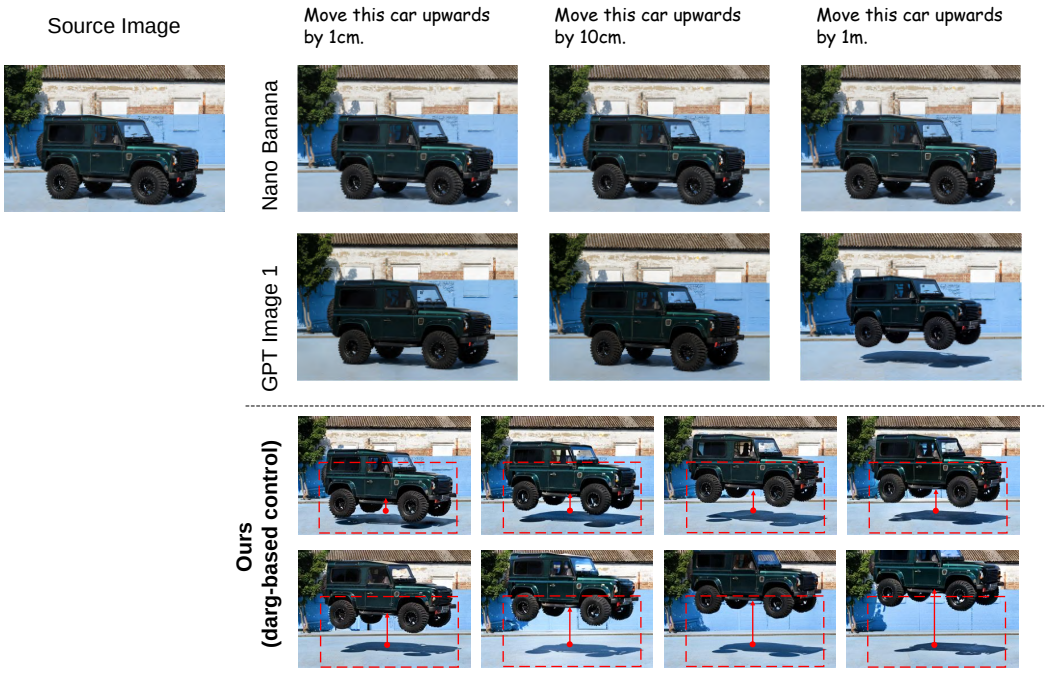

Figure 25: Qualitative comparison for subtle displacement editing.

E.5 Drag-based Image Editing

The current and next segmentation prediction tasks introduced in Sec. 3.3 not only support progressive planning and generation, but also enable controllable editing for enhanced user interaction. One such application is drag-based image editing for object displacement, scaling, and rotation, as illustrated in Fig.24. In this setting, users first provide an editing prompt to localize the RoE. Then, drag operations are applied to perform geometric transformations of the RoE. The transformed segmentation map driven by the transformation is incorporated into the context, allowing the model to generate a target image that adheres to the specified edits.

Despite the strong understanding capabilities of VLMs, they may still struggle to detect subtle semantic or visual differences when two frames differ only minimally. As illustrated in Fig. 25, we first present qualitative results from the most advanced proprietary systems—GPT Image 1 (OpenAI, 2025a) and Nano Banana (DeepMind & Gemini, 2025)—on the task of subtle displacement editing. We then showcase our drag-based editing results, demonstrating that this challenging requirement can be effectively addressed through the more flexible and fine-grained control (*i.e.*, drag). This comparison highlights the versatility of our method.

F Limitations

Discussion of Other Potential Limitations. First, we use T5 to encode text, which restricts the model's ability to comprehend complex instructions and generate nuanced textual outputs. Integrating a vision-language model (VLM) into the framework could significantly improve this capability. Second, while our framework demonstrates preliminary but promising emerging abilities, these can be further enhanced through supervised fine-tuning (SFT) on high-quality, application-specific datasets. Lastly, due to the high cost of querying VLM, we annotated only 10M training samples. Expanding both the model size and the dataset scale presents an exciting avenue for future research.

G Future Work

In the future, we aim to solve more challenging image creation tasks (Qu et al., 2023; Yang et al., 2024a; Qu et al., 2024) with complex and compositional prompts, by exploring multimodal chain-of-thought. Besides, post-training (Qu et al., 2025b; Gong et al., 2025) would stimulate more potential interesting abilities endowed by learning from videos. Finally, by introducing retrieved images (Qu et al., 2025a; Chen et al., 2022; Qu et al., 2021) into context, our model could achieve knowledge-intensive visual creation scenarios via retrieval-augmented generation.

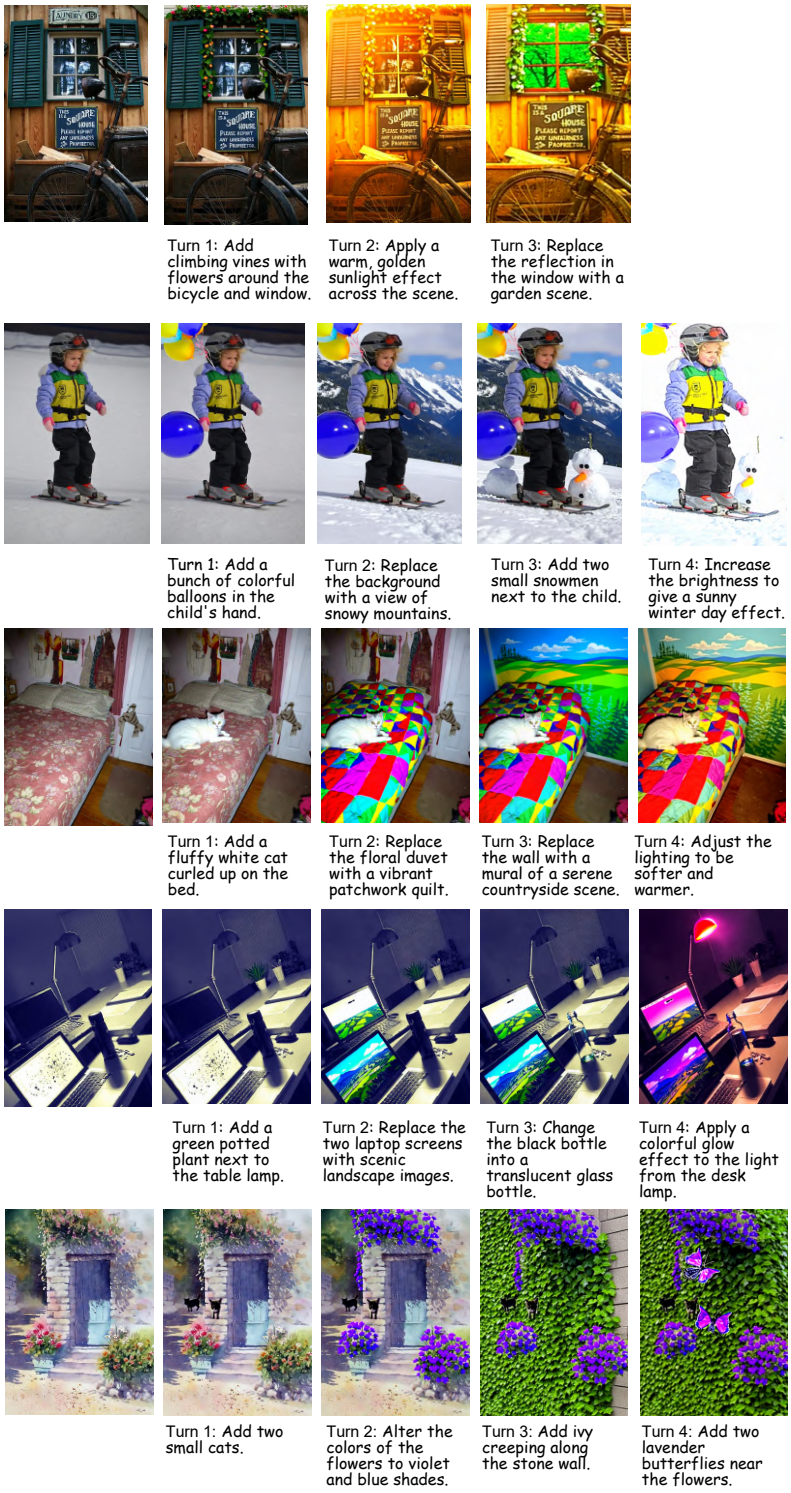

Figure 26: More qualitative results (1/4) of multi-turn image editing achieved by our method.

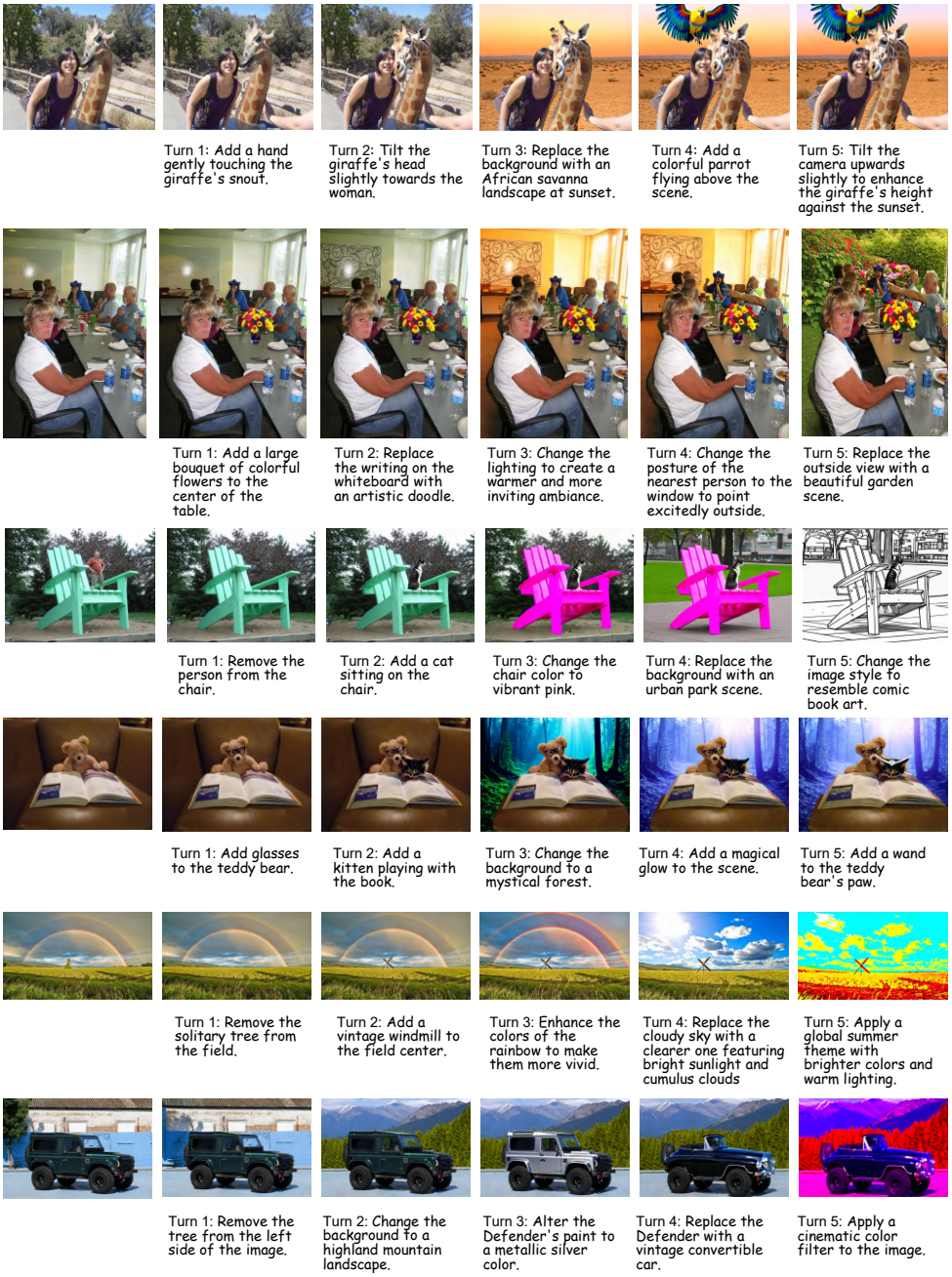

Figure 27: More qualitative results (2/4) of multi-turn image editing achieved by our method.

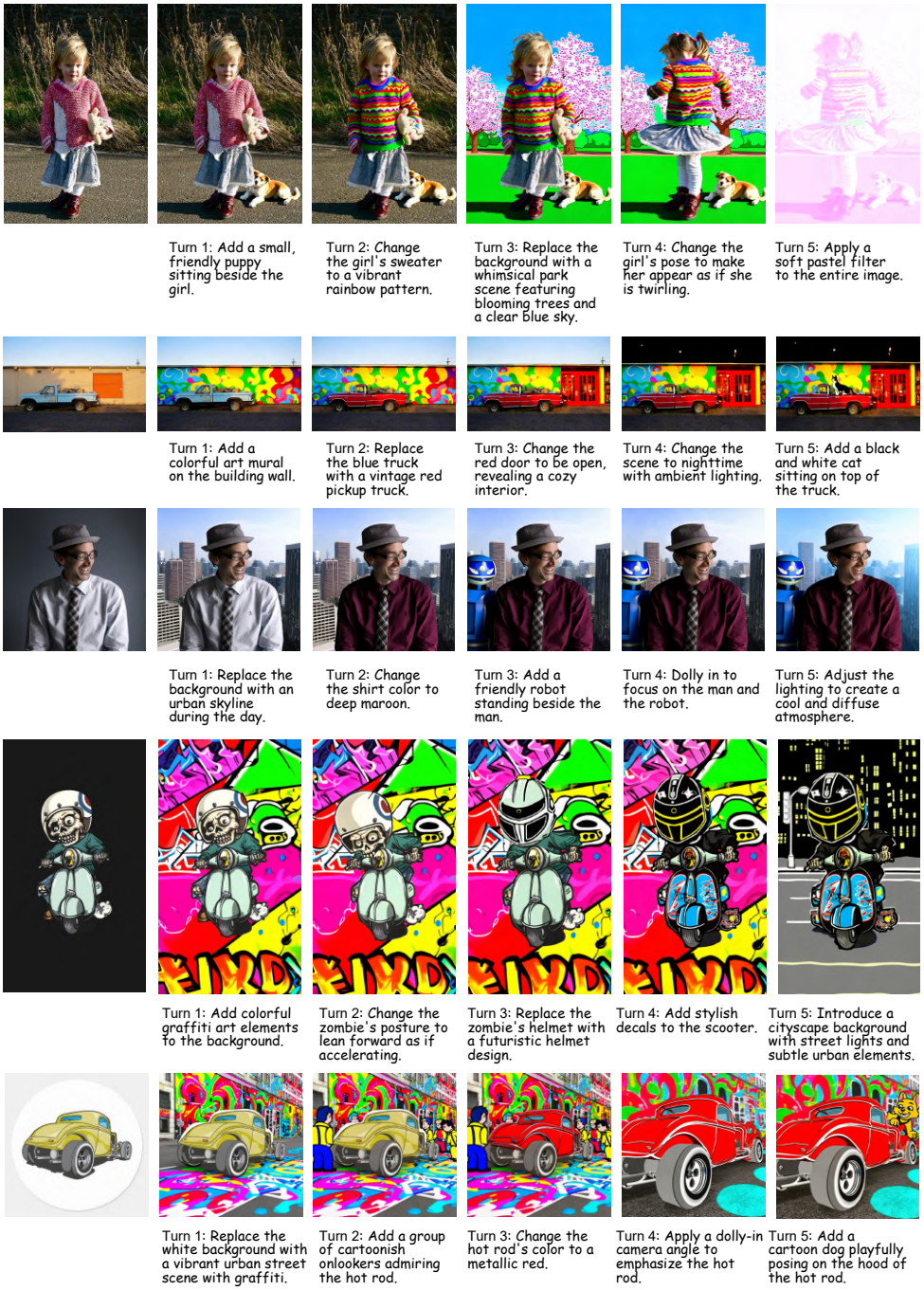

Figure 28: More qualitative results (3/4) of multi-turn image editing achieved by our method.

Figure 29: More qualitative results (4/4) of multi-turn image editing achieved by our method.