iMontage: Unified, Versatile, Highly Dynamic Many-to-many Image Generation

Zhoujie Fu[1,2], Xianfang Zeng[2,‡], Jinghong Lan[2], Xinyao Liao[1,2], Cheng Chen[1], Junyi Chen[3], Jiacheng Wei[1], Wei Cheng[2], Shiyu Liu[2], Yunuo Chen[2,3], Gang Yu[†,2] Guosheng Lin[†,1]

[1]Nanyang Technological University [2]StepFun [3]Shanghai Jiao Tong University

https://kr1sjfu.github.io/iMontage-web/

Figure 1. iMontage can flexibly deal with many input images, and can generate many output images with highly consistency. We use three different colors to represent three settings. The dotted-line box images are the input.

Abstract

Pre-trained video models learn powerful priors for generating high-quality, temporally coherent content. While these models excel at temporal coherence, their dynamics are often constrained by the continuous nature of their training data. We hypothesize that by injecting the rich and unconstrained content diversity from image data into this coherent temporal framework, we can generate image sets that feature both natural transitions and a far more expansive dynamic range. To this end, we introduce iMontage, a unified framework designed to repurpose a powerful video model into an all-in-one image generator. The framework

‡ Project Leader † Corresponding Author

consumes and produces variable-length image sets, unifying a wide array of image generation and editing tasks. To achieve this, we propose an elegant and minimally invasive adaptation strategy, complemented by a tailored data curation process and training paradigm. This approach allows the model to acquire broad image manipulation capabilities without corrupting its invaluable original motion priors. iMontage excels across several mainstream many-in-many-out tasks, not only maintaining strong cross-image contextual consistency but also generating scenes with extraordinary dynamics that surpass conventional scopes. Our code and model weights will be made publicly available.

1. Introduction

Large-scale diffusion-based generative models[16, 26, 38, 39, 44, 46] have sparked a revolution in creative and high-quality image generation, accelerating progress in downstream tasks such as image editing. A recent trend in the field is to unify diverse image tasks within a single framework [33, 35, 40, 59], inspired by the success of Large Language Models (LLMs) and large vision language models (VLMs) [12, 14, 36, 60]. While most unified image models remain specialized for single-in-single-out image tasks, certain commercial model has taken an early lead in extending unified image generation to multi-input, multi-output settings[3] very recently. Accordingly, the many-to-many (multi-input, multi-output) setting warrants systematic exploration by the academic and open-source communities.

The many-to-many paradigm splits into two approaches: (i) token-centric models that represent text and images as a unified multimodal token stream and autoregressively generate target tokens conditioned on the inputs[59], thereby achieving many-to-many mappings; and (ii) video-centric pipelines that repurpose video diffusion generation as backbones, casting the task as discontinuous video generation and naturally accommodating variable numbers of input and output frames[10, 30]. While the first approach provides an appealing and promising modal-unified solution, it's generation quality and instruction following capability is challenged by common sense compared to diffusion paradigm. In contrast, the second approach elegantly leverages pre-trained motion priors to markedly enhance temporal coherence and handle variable-length inputs and outputs. Specifically, [10] trained a model of video generation from scratch and constructed a large-scale dataset of captioned frame pairs for instruction-tuned editing, demonstrating strong consistency and faithful detail preservation with respect to the input images.

Despite these advances for the diffusion-based paradigm, a critical question persists: **How can a model generate highly dynamic multi-image outputs while maintaining temporal and semantic consistency?** To our empirical knowledge, image-only models can produce highly diverse images based on the same inputs, yet they struggle with temporal consistency due to limited implicit understanding of world dynamics. Meanwhile, video-based models bring strong motion priors that improve temporal consistency; however, most foundation video models are trained predominantly on contiguous clips, which rarely contains hard cuts, abrupt transitions, or large camera/subject motions, and thereby transferring poorly to highly dynamic content and limiting task versatility.

In response, we present **iMontage**, a unified generative model that produces multiple, highly dynamic images conditioned on instructions and arbitrary reference images. Following the video-based paradigm, iMontage builds on a large pretrained video model and treats both inputs and outputs as pseudo-frames. We introduce a novel rotary positional embedding (RoPE) strategy to prevent conceptual ambiguity between multiple image frames and video frames. Our strategy explicitly maintains the model's pretrained capability in modeling temporal coherence, while clearly differentiating the discrete nature of image sets from the continuous flow of video sequences. We further provide a data-curation pipeline, which is carefully categorized and filtered for motion diversity and instruction quality, supporting broad, highly dynamic scenario. Finally, we detail a training regimen that offers practical insights into multi-task unification. Together, these components marry video generation with many-to-many image generation, achieving both temporal and content consistency.

We evaluate iMontage across three settings: one-to-one image editing, many-to-one image generation, and many-to-many image generation. For each setting, we present strong qualitative performances across all sub-tasks, showcasing robust instruction following, high-dynamic outputs, and consistent content generation, as presented in Fig. 1. Furthermore, we provide state-of-the art quantitative metrics on image editing benchmark (one-to-one), in-context learning benchmark (many-to-one) and storyboard generation evaluation (many-to-many).

In summary, our contributions are as follows:

- We introduce **iMontage**, a unified model that handles variable numbers of input and output frames, bridging video generation and highly dynamic image generation.
- We develop a *task-agnostic, temporally diverse* data curation pipeline paired with a multi-task training paradigm, ensuring learnability across heterogeneous tasks and temporal structures and enabling robust many-to-many generalization.
- Our model showcases convincing results over huge number of variable experiments, including most mainstreaming image generation and editing tasks. Massive visualization results and comprehensive evaluation metrics provide SOTA results in open-source community and even comparable results with commercial models.

Figure 2. **Overview of iMontage.** The model accepts a flexible set of reference images and produces **N** outputs conditioned on a text prompt. Images are encoded by a 3D VAE separately, text by a language model, and both token streams are processed by an MMDiT. We concatenate clean reference-image tokens with noisy target tokens before denoising. *Right:* training uses fixed-length text tokens and variable-length image/noise tokens, transitions from dual stream to single stream blocks. For image branch, we apply *Marginal RoPE*, a head–tail temporal indexing that separates input and output pseudo-frames, preserves spatial RoPE, and supports many-to-many generation. In figure, notation H and W with subscription denote the height/width indices of the 2D RoPE computed at the image's native resolution, while notation T represents assigned time index for temporal dimension.

2. Related work

2.1. Unified Generation and Editing Models

Recent research has increasingly focused on consolidating diverse visual synthesis tasks into single, unified frameworks. Early efforts[17, 28, 58] like OmniGen [59] and ACE++ [35] pioneered monolithic architectures capable of handling generation, editing, and other vision tasks without requiring task-specific modules. This trend evolved with the integration of powerful multimodal large language models (MLLMs) as reasoning engines. Models such as Step1X-Edit [33] and Qwen-Image [14] leverage an MLLM to interpret complex user instructions, which then guide a diffusion decoder to produce high-fidelity edits. This approach significantly improves instruction-following capabilities. Notably, unified systems in other AIGC area are also emerging, such as [2, 34, 53] in video generation, [32, 50, 63] in audio generation, and even more powerful combining different modalities together[8, 13, 37, 47, 48]. While these unified image models demonstrate impressive versatility, they are predominantly architected for single-input, single-output tasks. They lack the inherent capability to manage multiple image inputs and generate a set of dynamically varied yet coherent outputs from a single prompt, a key limitation our work addresses.

2.2. From One-to-one to Many-to-many Generation

The frontier of generative modeling is advancing from single-image tasks to more complex many-to-many scenarios that require handling multiple inputs to produce multiple outputs. A significant paradigm shift was introduced by UniReal[10] , which re-frames multi-image generation as "discontinuous video generation." By leveraging the powerful temporal priors of video models, this approach naturally accommodates a variable number of input and output "frames" and uses large-scale video data as a source of universal supervision for learning real-world dynamics. Following this direction, models like RealGeneral[30] and Frame2Frame[45] also explore video backbones for unified image generation. More recent "any-to-any" models, such as BAGEL[60] and OmniGen[59], are trained on vast, interleaved multimodal datasets, enabling them to handle arbitrary combinations of inputs and outputs and exhibit emergent world-modeling capabilities. However, a critical challenge persists. Foundation video models are typically trained on contiguous video clips, which limits their ability to generate highly dynamic or temporally discontinuous content. This reliance on smooth motion priors hinders their versatility for tasks requiring abrupt scene changes or significant variations between outputs, a gap that iMontage is designed to fill.

3. Method

3.1. Model Design

Network Archtecture. As illustrated in Fig. 2, we adopt a hybrid-to-single-stream MMDiT paired with a 3D VAE for images and a language model for text instruction. All

Figure 3. **Overview of our dataset:** Our dataset is constructed from four sources and is organized into two stages, comprising high-quality foundational data and multiple task-oriented subsets.

components are initialized from HunyuanVideo [24]: the MMDiT and 3D VAE are taken from the I2V checkpoint, while the text encoder is taken from the T2V checkpoint. Reference images are encoded by the 3D VAE seperately and then patchified into tokens; textual instructions are encoded by the language model into text tokens. Following the I2V formulation, we concatenate clean reference-image tokens with noisy target tokens and feed the sequence to the image branch block. We train our model to accommodate variable numbers of input and output frames by constructing variable-length attention maps over their image tokens, guided by prompt-engineering cues. During training, we frozen VAE and text encoder, only full-finetune the MMDiT.

Position Embedding A key objective is to endow the transformer with sensitivity to multiple images without perturbing its original positional geometry. We adopt a simple yet effective strategy: cast all input/output images as pseudo-frames along the temporal axis, assign each a unique time index, and keep their native spatial resolution and 2D positional encoding intact. Concretely, we preserve the pretrained spatial RoPE and introduce a separable temporal RoPE with per-image index offsets, supplying cross-image ordering cues while leaving the spatial distribution unchanged. Inspired by L-RoPE [25], we assign input images to early temporal positions and output images to late positions. In practice, we allocate a 3D RoPE with 32 temporal indices, reserving $\{0, \ldots, 7\}$ for inputs and $\{24, \ldots, 31\}$ for outputs, leaving a wide temporal margin between them. This head–tail layout reduces positional interference between inputs and targets and empirically promotes more diverse output content while preserving temporal coherence.

Prompt Engineering We adopt a purely text-instruction interface powered by a strong LLM encoder, without masks or auxiliary visual embeddings. To unify heterogeneous tasks, we pair a set of common prompts with task-specific templates. For the common prompts, we (i) prepend a system-style preamble: *Please output N images according to the instruction:* and (ii) use an interleaved multimodal format that explicitly marks image positions via textual placeholders `<image_n>` within the prompt.

3.2. Dataset Creation

We divide our data construction into two phases: a pretraining dataset and a supervised fine-tuning (SFT) dataset. The overview of our dataset construction is refered in Fig. 3.

3.2.1. Pretraining Dataset

We partition the pretraining data into two pools: an *image-edit* pool and a *video frame-pair* pool, sourced from internal corpora. The image-edit pool spans most single-image editing tasks, providing paired (input, edited) images with concise, fine-grained instructions specifying the operation. The video frame-pair pool consists of high-quality frame pairs extracted from videos (with associated captions), curated under stringent quality criteria. We further refine the video

Table 1. Comparison metrics of **Motion Change** and **Edit overall** on GEdit-GPT4o-EN; **Action** and **Average** on ImgEdit. For GEdit-GPT4o-EN, Semantic Consistency (G_SC), Perceptual Quality (G_PQ), and Overall Score (G_O) are reported. **Bold** means the best performance and underline means the second best performance.

Category	Models	Motion Change - GEdit			Edit overall - GEdit			ImgEdit	
		$G_SC\uparrow$	$G_PQ\uparrow$	$G_O\uparrow$	$G_SC\uparrow$	$G_PQ\uparrow$	$G_O\uparrow$	**Action**	**Average**
Closed-source	Gemini 2.5[12]	6.87	7.79	6.72	8.25	8.29	7.89	4.61	4.30
	GPT-4o[36]	7.81	8.53	7.81	8.74	7.67	8.01	4.83	4.30
	Seedream 4.0[3]	5.58	8.53	5.53	8.41	8.04	7.81	4.66	4.32
Open-source	ICEdit[70]	0.93	7.98	1.13	4.94	7.39	4.87	3.68	3.05
	Omnigen[59]	3.35	6.68	3.12	5.88	5.87	5.01	3.38	2.96
	Omnigen2[54]	4.75	8.08	5.13	7.16	6.77	6.41	**4.68**	3.44
	Bagel[14]	5.25	8.03	5.09	7.48	6.80	6.60	4.17	3.20
	UniWorld-V1[29]	1.58	7.55	1.76	4.93	7.43	4.85	2.74	3.26
	HiDream-I1 (E1)[4]	1.58	7.23	1.66	5.66	6.06	5.01	3.33	3.17
	HiDream-E1.1[20]	**5.55**	7.80	**5.64**	7.15	6.65	6.42	4.18	3.97
	Flux-Kontext-dev[26]	5.23	7.53	4.95	7.16	7.37	6.51	4.35	3.97
	Step1X-Edit v1.1[33]	4.65	8.15	4.73	**7.66**	7.35	**6.97**	3.73	3.90
Open-source	**iMontage (Ours)**	5.25	**8.43**	5.53	7.21	**7.80**	6.94	4.48	**4.11**

frame pairs by selecting samples that satisfy the following filtering criteria:

For frame pairs drawn from a single clip, we apply motion filtering with an optical-flow estimator [49]: for each sample, we compute the average motion magnitude and preferentially retain or upweight high-motion instances to increase their prevalence. To further diversify dynamics, we concatenate segments from the same source video and re-clip them without motion- or camera-change heuristics (i.e., not cutting at large motions or pans), thereby producing cross transition frame pairs and mitigating the bias toward quasi-static content.

Post-filtering, the dataset comprises 5M image-edit pairs and 15M video frame pairs, providing supervision for highly dynamic content generating and robust instruction following.

3.2.2. Multi Task Dataset

Our Multi Task dataset is constructed based on tasks, varying from one-to-one task to many-to-many task. Our data curation pipeline for each task is described as follows:

Multi CRef. We crawl web posts to assemble reference images for human, object, and scenario. Human images are filtered to single-person shots via a detector [15]; object/scenario images need no extra filtering. A VLM [1] composes CRef prompts by randomly combining sources, GPT-4o [36] generates the corresponding images, and the VLM then scores and filters candidates. This pipeline yields around 90k high-quality samples.

Conditioned CRef. Different from the CRef dataset, we collect the data from an open-source dataset Echo-4o[67]. We apply some classic ControlNet[69] generation control maps to the target image. We use OpenPose[5] to generate the character poses of the composite image, use Depth-Anything-V2[64] to generate the depth map of the target image, and also use the Lineart model[22] as an edge detector. We add these condition pairs to Echo-4o and create a new Conditioned CRef dataset about 50k samples.

SRef. We curate style-reference data analogously to CRef. We scrape character posts and select human images via a VLM aesthetic score [1] as *content* references, and collect hand-drawn illustrations from open sources as *style* references. Using subject–style models[56, 61], we generate images by randomly pairing content and style. A VLM then scores outputs and checks ID consistency with the content image to prevent style leakage. This yields 35k samples.

Multi Turn Editing. In this task, we generate multiple responses at the same time according to instruction, where sub-steps instruction cover all editing tasks in pretraining image-edit dataset. Our data is extracted from an internal dataset and we collect around 100k samples.

Multi View Generation. We curate our multi-view dataset from the open-source 3D corpus MVImageNet V2 [19]. For each base sample, we randomly select 1–4 additional viewpoints and, in successive order, use GPT-4o [36] to caption the relative camera motion between adjacent images, yielding concise supervision for multi-view generation. We collect around 90k samples.

Storyboard Generation. Storyboard generation is closely related to the storytelling setting [42, 52], but targets high inter-panel diversity, for example, drastic scene changes and distinct character actions across images. Leveraging recent commercial foundation model Seedream4.0[3], we distill high-quality supervision from their outputs to construct instruction–image sequences for training. We begin with an internal character image dataset and apply a face-detection filter [15] and an NSFW filter [27] to obtain whole-face character reference images. We then design instruction templates that prompt Seedream4.0 to produce semantically rich, dynamic scenes and multi-panel stories. The generated images are captioned with GPT-4o [36], yielding con-

Table 2. Quantitative comparison on OmniContext grouped by model availability. "Char. + Obj." indicates Character + Object.

Category	Model	SINGLE		MULTIPLE			SCENE			Average↑
		Char.	Obj.	Char.	Obj.	Char. + Obj.	Char.	Obj.	Char. + Obj.	
Closed-source	Gemini 2.5[12]	8.62	9.11	8.77	8.88	7.39	7.29	7.05	6.68	7.84
	GPT-4o[36]	8.90	9.01	9.07	8.95	8.54	8.90	8.44	8.60	8.80
Open-source	InfiniteYou[23]	6.05	–	–	–	–	–	–	–	–
	OmniGen[59]	7.21	5.71	5.65	5.44	4.68	3.59	4.32	5.12	4.34
	UNO[57]	6.60	6.83	2.54	5.61	4.39	2.06	3.33	4.37	4.71
	BAGEL[14]	5.48	7.03	5.17	6.64	6.24	4.07	5.71	5.47	5.73
	OmniGen2[54]	**8.05**	7.58	**7.11**	7.13	7.45	6.38	6.71	7.04	7.18
Open-source	**iMontage (Ours)**	7.94	**7.77**	6.75	**7.57**	**8.20**	**6.90**	**6.81**	**7.37**	**7.41**

cise storyboard (instruction, images) pairs for supervision. We collect around 29k samples.

3.3. Training Scheme

We adopt a three-stage training strategy using a dynamic mixture of the curated data described above-specifically, a Pre-training stage for large-scale pre-training, a Supervised Fine-tuning stage, and a High-Quality Annealing stage:

Pre-training Stage. In this stage, we train on the *Pretraining Dataset* to instill instruction following and acclimate the model to highly dynamic content. Since we initialize from a pretrained backbone, we eschew progressive resolution schedules [7, 16, 18]; instead, we adopt aspect-ratio–aware resolution bucketing: for each sample, we select the best-fitting size from a set of 37 canonical resolutions and resize accordingly. Batch size in this stage is dynamically adjusted by sequence length, equalizing the token budget across resolutions and yielding smoother, more stable optimization.

SFT Stage. We investigate the best solution of unifying multitasks with huge variance in this stage. Our strategy can be concluded as follows:

- **FlatMix: All-in-One Joint Training.** Train all tasks together in a single mixed pool.
- **StageMix: Curriculum Training.** Two-phase schedule: first train on the three many-to-one tasks, then add the three many-out tasks and continue mixed training.
- **CocktailMix: Difficulty-Ordered Fine-Tuning.** We witness notable training difficulty variance for each single task, motivating us of a mixture of training by difficulty. In practice, we begin with the simplest task, then introduce the second-easiest while reducing the sampling weight of the first. We continue this process by adding one harder task at a time and gradually shifting mixture weights until the hardest task is included and receives the largest training share.

For the final decision, we choose the CocktailMix training strategy, and discussion about the training is detailed in the ablation study (Sec. 4.4). During all mixture training, we apply weights based on the data amount of each task, ensuring all tasks are treated equally. In this stage, we allow different resolution for input images while fix output resolution for convenience. Since input images can be different resolution, we set batch size per GPU to 1 during all SFT training stage.

HQ Stage. In image and video generation, it is widely observed that concluding training with a small tranche of high-quality data improves final fidelity [39, 65, 72]. We adopt this strategy: using a combination of manual review and VLM assistance, we curate high-quality subsets for each task, then perform a brief, unified finetuning pass across all tasks after SFT. During this stage, we anneal the learning rate to zero.

All our experiments all conducted on 64 NVIDIA H800 GPUs. We apply a constant learning rate of 1e-5 for all training stages and the training target follows flow matching[31]. More detailed implementation can be found in Sec. 6.

4. Experiment

As a unified model, iMontage shows strong performance on various tasks even compared to fixed input/output models. Note that our model only need one inference, with a default of 50 diffusion steps. For clarity, we organize results by input–output cardinality, spliting into one-to-one editing (Sec. 4.1), many-to-one generation (Sec. 4.2) and many-to-many generation (Sec. 4.3).

4.1. One-to-one Editing

We report competitive quantitative metrics and compelling qualitative results on instruction-based image editing. We compare our model against twelve strong baselines, including native image editing models, unified MLLM models and powerful closed-source product. Average metrics on GEdit benchmark[33] and ImgEdit benchmark[68] can be found in Tab. 1. Despite closed-source models and commercial models, iMontage shows strong performance on both benchmark over other models.

We also report metric about motion-related sub-task in Tab. 1. Our method demonstrates superior motion-aware editing, exhibiting strong temporal consistency and motion

Figure 4. Comparison with three baselines on storyboard generation setting. Single character and many characters samples are presented.

priors. These gains are expected: we inherit strong world dynamic knowledge from a large pretrained video backbone, then reinforce it with pretrained on a high-dynamics video–frame corpus. Please find our one-to-one image editing visualization results in Fig. 6 and Fig. 7.

4.2. Many-to-one Generation

The core challenge for multiple inputs is how to preserve all their content and harmony them together. We report our results on the OmniContext benchmark[54], which aims to provide a comprehensive evaluation of the models' in-context generation abilities. We report our metrics against seven baselines. Detailed metrics can be found in Tab. 2.

We also visualize representative results in supplementary materials, showing that iMontage handles diverse tasks while maintaining the source image's context. We select challenging cases to stress control and fidelity. In *Multi-CRef*, the model fuses cues from multiple references without altering core content, while being faithful to complex instruction by generating a highly detailed background. In *Conditioned CRef*, it respects the conditioning signal yet preserves the human's details, which is considered to be hard for generation models. For *SRef*, we include scene-centric and human/object-centric inputs to demonstrate strong style transfer that retains style and identity.

4.3. Many-to-many Generation

Generating multiple outputs while preserving consistency is highly challenging. We raise the bar by requiring both cross-output content consistency and temporal consistency.

To evaluate capability, we consider three disparate tasks. **Multi-view generation.** We simulate camera rotations, following [14], and use natural-language descriptions of camera motion to render novel views from a single reference image. This temporally continuous setting probes whether the model preserves identity, geometry, materials, and background context as the viewpoint changes. We report identity/structure consistency across views and visualize long arcs of rotation to stress continuity. All our visualization results can be found in Fig. 10.

Table 3. Storyboard generation metrics over iMontage (ours) and three baselines. Dino feature similarity, Clip feature similarity and VLM rating scores are reported.

(a) **Identity Preservation**.

Method	DINO↑	CLIP↑	VLM_{pref} ↑
StoryDiffusion	0.367	0.570	3.962
UNO (w/ UMO)	0.519	0.674	6.625
OmniGen2 (w/ UMO)	0.486	0.619	6.857
iMontage (Ours)	**0.585**	**0.690**	**7.909**

(b) **Temporal Consistency**.

Method	DINO↑	CLIP↑	VLM_{pref} ↑
StoryDiffusion	0.440	0.649	7.111
UNO (w/ UMO)	0.479	0.676	6.556
OmniGen2 (w/ UMO)	0.460	0.655	7.889
iMontage (Ours)	**0.615**	**0.745**	**9.556**

Multi-turn editing. Most image editors support multi-turn pipelines by running inference sequentially, yet they often drift, overwriting non-target content. We cast multi-turn editing as a content-preservation task: given an initial image and a sequence of edit instructions, the model should localize changes while maintaining other parts. All our visualization results can be found in Fig. 7.

Storyboard generation. This is our most comprehensive setting: temporally, the model must produce smooth, continuous trajectories while also handling highly dynamic transitions such as hard cuts, large camera or subject motions, and scene changes; spatially, it must preserve con-

Table 4. User study metrics on storyboard generation of twenty samples. Rating scores are between 1 and 5, while a higher score means better performance.

(a) Instruction following (IF) and identity preservation (IP).

Method	IF↑	IP↑
StoryDiffusion	2.81	1.86
UNO (w/ UMO)	3.68	2.90
OmniGen2 (w/ UMO)	3.97	3.07
iMontage (Ours)	**4.46**	**3.91**

(b) Temporal consistency (TC) and overall quality (OQ).

Method	TC↑	OQ↑
StoryDiffusion	2.28	2.12
UNO (w/ UMO)	3.05	3.03
OmniGen2 (w/ UMO)	3.04	3.23
iMontage (Ours)	**4.31**	**4.16**

tent consistency by maintaining identity, layout, and fine-grained appearance across all outputs.

As illustrated in visualization results in supplementary material, iMontage delivers coherent yet highly diverse results across all three settings in a *single* forward pass. To the best of our knowledge, this is the first model to unify these tasks within one model and one-shot inference.

To better quantify many-out capability, we conduct a quantitative study in the storyboard setting, comparing our method against two unified systems (OmniGen2 and UNO) and a storytelling-focused baseline, StoryDiffusion [73]. We focus on two axes: ID preservation and temporal consistency. The former measures how closely each generated character matches the reference identity (especially the character's whole body details, such as clothes, skin color, hair), while the latter captures cross-panel coherence among the generated images. In our evaluation, the evaluated OmniGen2 and UNO models are optimized by UMO[11], which improves identity preservation and other quality measures. For metrics, we use DINO[6] and CLIP[41] feature similarity following [21, 71], along with a VLM rating system. We report the comparison score in Tab. 3. We also present visualization comparison in Fig. 4. Detailed conduction of our storyboard evaluation can be found in Sec. 8.1.

Furthermore, for a more comprehensive evaluation, we conduct a user study with 50 professional participants. We show the comparison metrics in Tab. 4. Our method achieves the best performance both at instruction following and identity preservation, outperforms baselines with a big margin. Detailed experiments of user study can be found in Sec. 8.2.

4.4. Ablation Study

RoPE Strategy. We first ablate our RoPE strategy design. Our default *Marginal RoPE* assigns inputs to the head of the temporal index range and outputs to the tail, leaving a gap between them; the control, *Even RoPE*, distributes all images uniformly along the temporal axis. We conduct our ablation study using a same setting from pretraining dataset, of which is only a small amount of data. We observe a late convergence for Even RoPE, with the same training steps. Fig. 5 indicates the visualization of the RoPE ablation study.

Figure 5. **Ablation on different RoPE strategy.** We evaluate on a subset of the editing data with low resolution, training each strategy for the same number of steps. In the figure, corner numbers indicate provenance: **1** original input, **2** edited ground truth, **3** output from *Marginal RoPE*, and **4** output from *Even RoPE*.

Training Scheme. As discussed in Sec. 3.3, we ablate three SFT strategies. With *FlatMix*, the training loss oscillates strongly and does not stabilize. After some updates, the model drifts toward the easier tasks even with inverse-size reweighting. We conduct *StageMix* and *CocktailMix* experiments at the same time, the former groups training by task type, while the latter organizes the schedule by task difficulty. CocktailMix delivers strong results across all tasks and shows a clear advantage on the harder settings, outperforming StageMix by a significant margin. We also conduct a comparison experiment on Multi CRef, with a same training steps for both strategy. The result reveals a 12.6% gain on OmniContext for CocktailMix. We show more details in Sec. 8.3.

5. Conclusion and Limitations

In conclusion, we introduce iMontage, a unified many-to-many image generation model that can create highly dynamic contents while preserving both temporal and content consistency. Adequate experiments demonstrate iMontage's superior capabilities in image generation.

However, iMontage still face some limitations. First, due to data and compute constraints, we have not explored long-context many-to-many settings, and the model currently delivers its best quality with up to four inputs and four outputs. Second, several capabilities remain limited. We provide a detailed breakdown and failure cases in Sec. 9.2. We also include more discussion about concurrent work in Sec. 9.1. For next step, we view scaling long-context supervision, enhancing data quality and broadening task coverage as primary directions for future work.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[3] Bytedance. Seedream4.0, 2025. https://seed.bytedance.com/en/seedream4_0. 2, 5

[4] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. 5

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 5

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 8, 1

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 6

[8] Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025. 3

[9] Lan Chen, Yuchao Gu, and Qi Mao. Univid: Unifying vision tasks with pre-trained video generation models. *arXiv preprint arXiv:2509.21760*, 2025. 3

[10] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. 2, 3

[11] Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025. 8

[12] Google Deepmind. Gemini2.5, 2025. https://deepmind.google/models/gemini/pro/. 2, 5, 6

[13] Google Deepmind. Veo3, 2025. https://deepmind.google/models/veo/. 3

[14] 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. 2, 3, 5, 6, 7

[15] Arnab Dhar. Yolov8-face-detection, 2024. https://huggingface.co/arnabdhar/YOLOv8-Face-Detection. 5

[16] 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. 2, 6

[17] Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing. *arXiv preprint arXiv:2503.12652*, 2025. 3

[18] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 6

[19] Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. Mvimgnet2. 0: A larger-scale dataset of multi-view images. *arXiv preprint arXiv:2412.01430*, 2024. 5

[20] HiDream-ai. Hidream-e1-1, 2025. https://huggingface.co/HiDream-ai/HiDream-E1-1. 5

[21] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 8, 1

[22] Huggingface. Controlnet auxiliary models. https://github.com/huggingface/controlnet_aux?tab=readme-ov-file, 2023. 5

[23] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025. 6

[24] 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. 4, 3

[25] Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025. 4

[26] Black Forest Labs. Flux.1 [dev], 2024. https://huggingface.co/black-forest-labs/FLUX.1-dev. 2, 5

[27] LAION. Clip-based nsfw detector, 2021. https://github.com/LAION-AI/CLIP-based-NSFW-Detector. 5

[28] Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. *arXiv preprint arXiv:2401.13388*, 2024. 3

[29] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 5

[30] 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. 2, 3

[31] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 6

[32] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2871–2883, 2024. 3

[33] 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. 2, 3, 5, 6, 1

[34] Jiabin Luo, Junhui Lin, Zeyu Zhang, Biao Wu, Meng Fang, Ling Chen, and Hao Tang. Univid: The open-source unified video model. *arXiv preprint arXiv:2509.24200*, 2025. 3

[35] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025. 2, 3

[36] OpenAI. Gpt4o, 2024. https://www.openai.com/. 2, 5, 6, 1

[37] OpenAI. Sora2, 2025. https://openai.com/index/sora-2/. 3

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[39] 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. 2, 6

[40] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025. 2

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 8, 1

[42] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2493–2502, 2023. 5

[43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1

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

[45] Noam Rotstein, Gal Yona, Daniel Silver, Roy Velich, David Bensaid, and Ron Kimmel. Pathways on the image manifold: Image editing via video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7857–7866, 2025. 3

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

[47] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025. 3

[48] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025. 3

[49] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 5

[50] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023. 3

[51] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3

[52] Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human efforts. *International Journal of Computer Vision*, pages 1–22, 2024. 5

[53] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025. 3

[54] 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*, 2025. 5, 6, 7, 1, 3

[55] Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Tianshi Cao, Kai He, Yifan Lu, Ruiyuan Gao, Enze Xie, Shiyi Lan, Jose M Alvarez, et al. Chronoedit: Towards temporal reasoning for image editing and world simulation. *arXiv preprint arXiv:2510.04290*, 2025. 3

[56] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025. 5, 3

[57] 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*, 2025. 6

[58] Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28533–28543, 2025. 3

[59] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 2, 3, 5, 6

[60] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2, 3

[61] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 5, 3

[62] Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. Withanyone: Towards controllable and id consistent image generation. *arXiv preprint arXiv:2510.14975*, 2025. 3

[63] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023. 3

[64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 5

[65] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 6

[66] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[67] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025. 5

[68] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 6

[69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 5

[70] 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*, 2025. 5

[71] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 8, 1

[72] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 6

[73] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37: 110315–110340, 2024. 8

iMontage: Unified, Versatile, Highly Dynamic Many-to-many Image Generation

Supplementary Material

6. Implementation Details

For training, we treat all DiT blocks as trainable components in all stages, frozen VAE and text encoders. In pretraining stage, we start with more video clip data and less image editing data, then gradually counting more image editing data for better instruction following capability. The ratio is a linear increase of 25% to 75% for image editing data. In this stage, we adopt dynamic resolution grouping. We choose 512^2, 768^2 and 1024^2 as base buckets and derive ±32-pixel variants on both height and width, yielding candidate resolution of 37 categories. Each training image is assigned to the candidate that best matches its native size (preserving aspect ratio via short-side resize and optional padding), and then resized accordingly. Batch size for 512-resolution bucket per gpu is 8, while 4 for 768 bucket and 2 for 1024 bucket. In the SFT stage, due to data constraints, each task is trained at a fixed resolution. At inference, however, the model generalizes well to arbitrary resolutions across tasks, exhibiting stable behavior and consistent quality without task-specific resizing rules. In the HQ stage, we collect a set of high-aesthetic, higher-resolution multi-task samples and mix them with curated subsets from earlier datasets, then perform an annealed finetuning pass. All our training are conducted on NVIDIA H800 gpus, takes about 7 days to cover all training stage on 64 H800 gpus. More detailed hyperparameters used in training stage can be found in Tab. 5.

For inference, we conduct classifier-free guidance (CFG) for text embeddings. The default cfg is 6.0 for all inference tasks, and inference steps is set to 50. To align unconditional generation, we adopt a 0.1 probability for none caption in all training stage.

7. More Qualitative Results

We present more visualization results to reveal the powerful capability of our model. Please find our image editing results in Fig. 6 and Fig. 7, multi cref results in Fig. 8 and multi view results in Fig. 10.

8. Detailed Experimental Details

8.1. Storyboard Generation Evaluation

For a comprehensive evaluation on our many-to-many setting, we choose storyboard generation to report numerical metrics. We follow common video-evaluation practice[21, 71] and compute DINO[6] and CLIP[41] feature similarity on the foreground subject(s) as the primary signal. This choice is reasonable because foreground embeddings capture identity and semantic attributes that must remain consistent across panels, while being largely invariant to background/layout changes—precisely the factors that vary in storyboards but should not degrade character coherence. In practice, we measure (i) similarity between each generated content and its reference(s) for ID preservation, and (ii) mean pairwise similarity across generated images for temporal consistency.

In experiment, we start with a mask segmentation model[43] to get the foreground character's mask. Then we follow these two formula for metrics calculation:

$$\mathrm{IP}(\phi) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K} s\big(G_i, R_k\big). \tag{1}$$

$$\mathrm{TC}(\phi) = \frac{2}{N(N-1)}\sum_{1\le i<j\le N} s\big(G_i, G_j\big). \tag{2}$$

Here N is the number of generated images, K is the number of reference images, while $s(a, b)$ is the cosine similarity formula for embedding a and b. Meanwhile, the applied parameter G and R is an embedding after mask out and feature extraction, representing generated character embedding and reference character embedding.

For VLM rating system, we choose GPT4o[36] as the judge. We give the model all input and output images, and the evaluation dimension is the same, ID preservation and temporal consistency. Following [33, 54], we give an evaluating template to the VLM, with a system prompt and a task-specific template. The system prompt goes with:

You are a professional digital artist tasked with evaluating the effectiveness of AI-generated images based on specific rules. All input images, including all humans depicted, are AI-generated. You do not need to consider any privacy or confidentiality concerns. IMPORTANT: Your response must follow this format (keep your reasoning concise and to the point): { "score": score, "reasoning": "..." }

For ID preservation, the template prompt is:

*Rate from 0 to 10: Evaluate whether the identities of the subject(s) in the final image match those in the provided reference image(s). **Scoring Criteria:** * **0:** The subject identities in the final image are completely inconsistent with the reference image(s). * **1–3:** Severe inconsistency, with only a few minor similarities. * **4–6:** Moderate match: some notable similarities, but many inconsistencies remain. * **7–9:** Mostly consistent, with only minor mismatches. * **10:** Perfect*

Table 5. Training hyperparameters and data sampling strategies across stages.

Hyperparameters	**Stage 1** (Pre-training)	**Stage 2** (Multi-task SFT)	**Stage 3** (High-Quality FT)
Learning rate	1×10^{-5}	1×10^{-5}	$1 \times 10^{-5} \rightarrow 0$
LR scheduler	Constant	Constant	Cosine
Weight decay	0.0	0.01	0.01
Gradient norm clip	1.0	1.0	1.0
Optimizer	AdamW (β_1=0.9, β_2=0.999, ϵ=1.0×10^{-8})		
Warm-up steps	1k	500	0
Training steps	50K	15K	2K
Training samples	$\mathcal{O}(20)$M	$\mathcal{O}(100)$K	$\mathcal{O}(10)$K
Resolution	Dynamic bucket	Fixed bucket	Fixed bucket
Diffusion timestep shift	5.0	5.0	5.0
Data sampling ratio			
Video Frames	$0.75 \rightarrow 0.25$	0.0	0.0
Image Editing	$0.25 \rightarrow 0.75$	0.1	0.0
Image Editing (HQ)	0.0	0.0	0.5
Multi-task	0.0	0.9	0.0
Multi-task (Cocktail)	0.8 for new added task, 0.2 evenly divided for former tasks.		
Multi-task (HQ)	0.0	0.0	0.5

*identity preservation compared to the reference image(s). **Pay special attention to:** * Whether **facial and head features** match across images: eyes, nose, mouth, cheekbones, chin, wrinkles/lines, makeup, hairstyle, hair color, overall facial structure and head shape. * **Body shape/proportions** and **skin tone** consistency; watch for abnormal anatomical changes. * **Clothing and accessories** if the instruction does not request changes; otherwise do not penalize expected edits. * Distinctive attributes (moles, scars, freckles, tattoos, piercings) that should persist. * If multiple references are given, ensure the correct individual(s) from each reference are present and not confused. **Do not** assess composition, pose, background, or aesthetics unrelated to identity preservation. **Scoring should be strict** — avoid giving high scores unless the identity match is clearly strong. Editing instruction: instruction.*

And for temporal consistency, the template prompt is:

*Rate from 0 to 10: Evaluate whether the identities of all subject(s) remain consistent across the provided generated images (sequence or set). **Scoring Criteria:** * **0:** Subjects are completely inconsistent across images (identity changes or swaps occur). * **1–3:** Severe inconsistency; frequent identity drift, swaps, or major attribute changes. * **4–6:** Moderate consistency; some notable similarities but multiple mismatches across images. * **7–9:** Mostly consistent identities with only minor mismatches. * **10:** Perfect temporal identity consistency across all images. **Pay special attention to:** * Stable **facial/head features** for the same subject across images (eyes, nose, mouth, facial structure, hairstyle/color). * Consistent **body shape** and **skin tone** for each individual across images. * **Clothing/accessories** stability unless the instruction implies changes; otherwise do not penalize expected edits. * For **multi-person scenes**, ensure each person maintains a consistent identity mapping across images (no A/B swapping). **Ignore** differences in pose, composition, viewpoint, background, or lighting that do not affect identity. **Scoring should be strict** — do not award high scores unless identity consistency is clear across all images. Editing instruction: instruction*

8.2. User Study

We invite 50 participants, who are familiar with image and video generative models, to engage in our evaluation on storyboard generation. We curate twenty evaluation samples by first searching some high quality human photos from website, then manually craft some storyboard caption based on them. For fairness, we include reference subjects spanning three racial groups (black, white and yellow) and two genders (female and male), and we vary prompts from simple to complex. Each testing sample provides one or two reference characters and requests generation of two to four storyboard images.

We request participants to rate for all results in the same sample. The rating system follows four criteria scored on a 5-point Likert scale (1=Poor, 5=Excellent): (i) *Instruction Following*—whether the images follow the prompt; (ii) *ID Preservation*—consistency with the reference character(s), emphasizing facial and fine attributes; (iii) *Temporal Con-*

sistency—whether the same character remains consistent across the generated panels; and (iv) *Overall Quality*—a holistic judgment beyond adherence and consistency. For each sample, all competing models are rated by the same participant to reduce between-rater variance, and model identities are anonymized and presentation order is randomized. We then report scores for each metric based on the mean rating. We provide a showcase of our rating system in Fig. 12.

For a fair comparison, we use each model's recommended inference settings. Specifically: StoryDiffusion at 768×768, classifier-free guidance (CFG)=5.0, 50 inference steps; UNO at 768×768, CFG=4.0, 25 steps; and OmniGen2 at 1024×1024, CFG=5.0, 50 steps. Our model follows setting as 1024x640 resolution, CFG=5.0, 50 steps. All experiments are conducted based on a random seed. Note that for a single sample, other models should be inferred several times with the same seed; iMontage uses one seed, outputting many results for one inference.

We present the visualization results of evaluation from Fig. 13 to Fig. 19.

8.3. Training Scheme Ablation

We ablate three scheduling strategies for SFT: *FlatMix* (all tasks jointly), *StageMix* (grouped by task type), and *CocktailMix* (difficulty-ordered curriculum). We begin with *FlatMix* and then transition to difficulty-aware scheduling.

Task difficulty gap. Under a shared setup (data, optimizer, steps), we observe a clear difficulty spread across tasks: the easiest task, *multi-editing*, and the hardest task, *storyboard generation*, differ by roughly 0.2 in training loss. This gap motivates difficulty-aware mixing.

StageMix vs. CocktailMix. We train *StageMix* with the same protocol used for our Stage 2 and Stage 3 runs and compare it head-to-head with *CocktailMix*. On OmniContext[54], *StageMix* underperforms by 12.6% relative to *CocktailMix*. Other tasks all have worser visualization results for *StageMix*. These observations indicate that difficulty-ordered mixing yields better optimization stability and stronger generalization, especially on the harder tasks.

9. More Discussion

9.1. Concurrent Works

Though we are not the first unified image generation model developed upon video models[10, 30], we consider iMontage as the first practical many-to-many system for open-source community. Likewise, two very recent efforts build image capabilities on top of video backbones. ChronoEdit[55] treats the input and edited outputs as the first and last frames of a short "video" and jointly denoises them with temporal-reasoning tokens, leveraging a pretrained video generator to improve physical plausibility and temporal coherence in edits. UniVid[9] explores a complementary route: it adapts a pretrained video DiT with lightweight SFT to a broad suite of vision tasks—both understanding and generation—by casting tasks as "visual sentences," thereby avoiding task-specific architectural changes and generalizing across modalities and data sources.

Our model focuses on another area, narrowing the gap between image and video generation by casting image synthesis as a unified many-to-many problem. We view this as a practical technical pathway and plan to extend it into a more capable, fully unified system.

9.2. Observed Failure Case

Our model still exhibits failure cases on certain tasks, as illustrated in Fig. 11. For **image editing**, the most salient issue is near-zero ability to render Chinese characters (Fig. 11a), largely inherited from the base backbone HunyuanVideo [24], which lacks robust text-rendering supervision. For **SRef**, our training data are distilled from other models [56, 61], which is suboptimal; we observe occasional background leakage, which is a known challenge in style-reference transfer. Finally, we note a **head-detail mismatch** in some generations. This limitation stems from data constraints—namely, insufficient training coverage of diverse, high-detail head/face depictions. Two complementary remedies are promising: (i) adopt human-centric identity modules by injecting face embeddings [51, 62, 66]; and (ii) expand coverage of high-quality, head-focused data to strengthen fine-grained facial detail preservation.

Figure 6. Visualization results for image editing. Zoom in to see more details.

Figure 7. Visualization results for image editing. Zoom in to see more details.

Figure 8. Visualization results for multi CRef. Zoom in to see more details.

Figure 9. Visualization results for conditioned CRef and SRef. Zoom in to see more details.

Figure 10. Visualization results for multi view generation, which can be divided to object-centric and scene-centric. Zoom in to see more details.

(a) Add Chinese characters. (b) Background leakage. (c) Bad performance at head details.

Figure 11. Representative failure case for certain task. Zoom in to see more details.

Your task. For each prompt with **1–2 reference characters**, evaluate the **2–4 storyboard images** produced by each model. Rate **every model** on the same example. Ignore watermarks/branding.

Scoring dimensions:

1. Instruction Following: - Do the images follow the text prompt?
2. ID Preservation: - Across the generated images, do the same characters and key elements remain consistent?
3. Temporal Consistency: - Across the generated images, do the same characters and key elements remain consistent?
4. Overall Quality: - Holistic visual quality and usability (clarity, naturalness, composition).

Scoring Rules (1–5, 1=Poor, 5=Excellent):

1. Instruction Following: 1: Largely mismatched; 2–3: Partially matched, key elements missing/wrong; 4–5: Largely/fully matched with correct details.
2. ID Preservation: 1: Clear mismatch; 2–3: Roughly similar but noticeable detail errors; 4–5: Clearly consistent and recognizable with stable fine details.
3. Temporal Consistency: 1: Strong drift across images; 2–3: Main subject consistent but several details vary; 4–5: Good consistency with stable details.
4. Overall Quality: 1: Low quality/unnatural; 2–3: Acceptable with flaws; 4–5: High quality, natural, well-formed.

Ref Char	Prompt	Model_1	Model_2	Model_3	Model_4
	1. A woman in a kimono walks through a lush garden path, holding a red parasol; 2. 2. She kneels down to gently reach out to a white cat on a leaf-covered path; 3. 3. The woman strolls along a sunlit garden path, her parasol casting a shadow behind her.				
	Instruct Following	Score:	Score:	Score:	Score:
	ID Preservation	Score:	Score:	Score:	Score:
	Temporal Consistency	Score:	Score:	Score:	Score:
	Overall Quality	Score:	Score:	Score:	Score:

Figure 12. User study template.

Figure 13. User study comparison visualization results. Zoom in to see more details.

Figure 14. User study comparison visualization results. Zoom in to see more details.

Figure 15. User study comparison visualization results. Zoom in to see more details.

Figure 16. User study comparison visualization results. Zoom in to see more details.

Figure 17. User study comparison visualization results. Zoom in to see more details.

Figure 18. User study comparison visualization results. Zoom in to see more details.

Figure 19. User study comparison visualization results. Zoom in to see more details.