DuetGen: Towards General Purpose Interleaved Multimodal Generation

Anonymous CVPR submission

Paper ID 18050

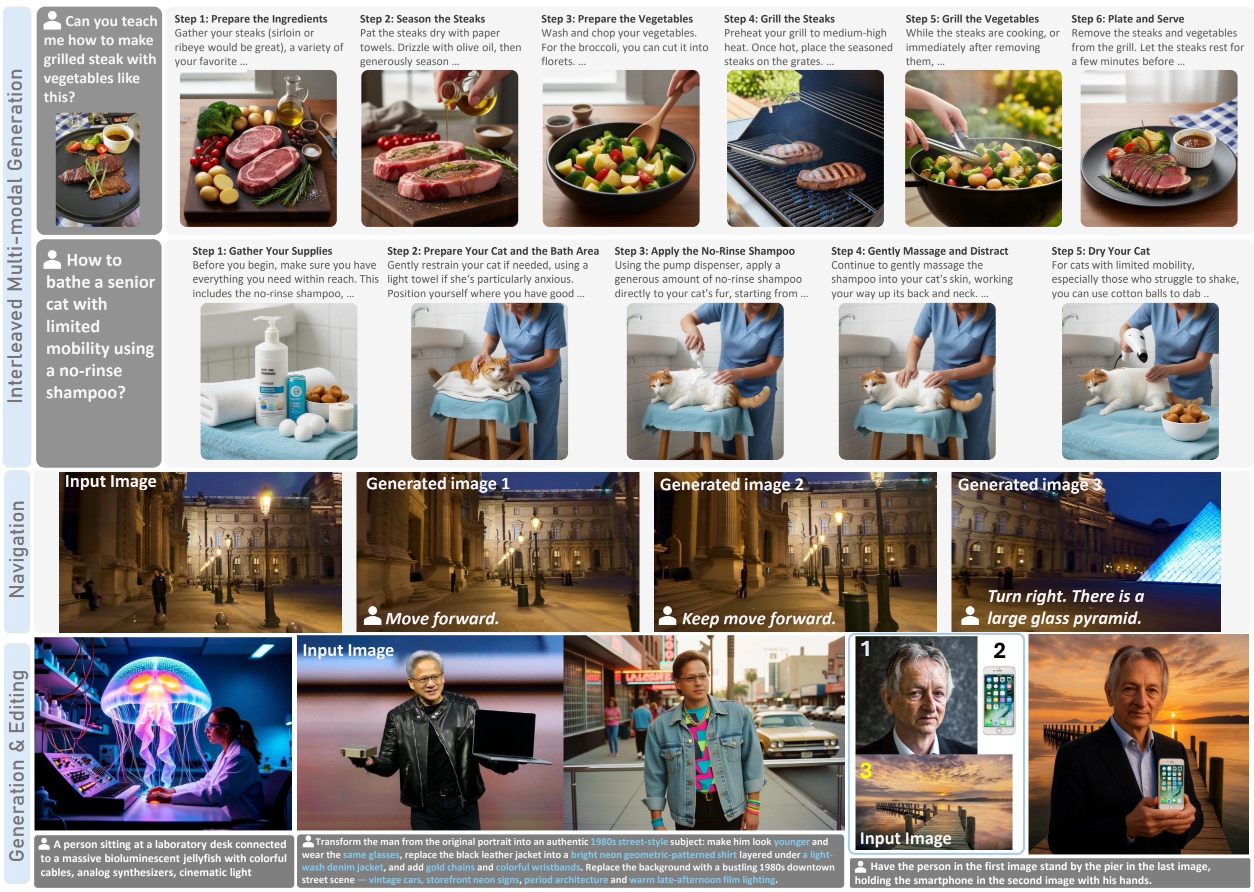

Figure 1. **Capabilities of DuetGen.** Beyond standard tasks like image understanding, generation, editing, and navigation, DuetGen supports interleaved multimodal content generation—a capability lacking in most unified models like Bagel [13].

Abstract

Interleaved multimodal generation enables capabilities beyond unimodal generation models, such as step-by-step instructional guides, visual planning, and generating visual drafts for reasoning. However, the quality of existing interleaved generation models under general instructions remains limited by insufficient training data and base model capacity. We present DuetGen, a general-purpose interleaved generation model that systematically addresses data curation, architecture design, and evaluation. On the data side, we build a large-scale, high-quality instruction-tuning dataset by combining multimodal conversations rewritten from curated raw websites, and diverse synthetic examples covering everyday scenarios. Architecturally, DuetGen leverages the strong visual understanding of a pretrained multimodal LLM and the visual generation capabilities of a diffusion transformer (DiT) pretrained on video generation, avoiding costly unimodal pretraining and enabling flexible base model selection. A two-stage decoupled strategy first instruction-tunes the MLLM, then aligns DiT with it using curated interleaved image–text sequences. Across public and newly proposed benchmarks, DuetGen outperforms prior open-source models in text quality, image fidelity, and

image–context alignment, and also achieves state-of-the-art performance on text-to-image and image editing among unified generation models. Code and data will be released.

1. Introduction

Interleaved text–image generation enables a critical class of applications requiring tightly coupled multimodal outputs—such as step-by-step instructional guides, visual planning, and interactive editing—where text and visuals must be produced in a coordinated manner. Although early works [16, 20, 53] show proof-of-concept results for storytelling or QA, they lack quantitative evaluation and are limited by their base models and data. Recent visual chain-of-thought systems [19,24,41,56] generate images as visual drafts interleaved with textual thinking, but only in limited domains such as math and navigation. Despite these efforts, the field still lacks a systematic approach to general-purpose interleaved generation, spanning data, training, and evaluation. To fill this gap, we present DuetGen, a framework that holistically addresses all three components.

A major bottleneck for interleaved generation is the lack of high-quality, diverse instruction-tuning data, especially data with realistic user–assistant interactions. Although instruction tuning is essential for (multimodal) LLMs [29, 36, 48], existing efforts largely rely on large-scale interleaved pretraining corpora [25, 63], or video dense captions [13]. These sources provide limited instruction-style supervision. Recent visual chain-of-thought studies [19, 24, 41] interleave images with text as visual drafts for tasks like geometry or navigation, but they target reasoning rather than high-quality interleaved generation, and their task coverage remains narrow. To address the quantity, quality, and diversity gaps in instruction-tuning data, we curate 298k interleaved conversation samples from two complementary sources: (1) a **data engine** that leverages a series of LLM/MLLM-based filtering and rewriting steps to convert raw webpages into clean user–assistant conversations; and (2) **synthetic data** generated by top-performing commercial models [18] using carefully curated prompts designed to elicit high-quality images. For the data engine, we scrape 347k webpages from how-to sites, filtering the invalid webpages and images, then rewrite and convert the remaining passages into 268k conversations. The MLLM+LLM pipeline improves linguistic quality, enforces image–text coherence, and enables generating user inputs in arbitrary multimodal formats. Though webpages provide coherent real-world descriptions, their image aesthetics and resolutions are often limited due to lack of quality control. To enhance visual quality, we leverage the current best commercial interleaved model, Nano Banana [18], to synthesize 30k high-quality interleaved samples. To ensure broad topic coverage, human annotators curate 1,500 seed prompts spanning 151 subcategories across 8 domains (*e.g.*, home & living, transportation), and we use OpenAI O3 [35] to expand them into a diverse prompt pool. This curated synthetic subset substantially improves the visual quality of our instruction-tuning data.

To establish basic interleaved generation abilities, most unified models [43, 53] adopt an early-fusion paradigm that jointly trains on interleaved and unimodal generation tasks, such as text and images. Some works [24, 41] also attempt to fine-tune from these pretrained models. However, unimodal pretraining requires heavy data engineering and computation, and restricts the choice of base models when scaling to different capacities. Recent unified systems [26, 37, 50] combine pretrained image generators with MLLMs, but their interleaved generation remains underexplored or limited by architectural constraints. For example, the adopted image generation heads cannot accept multiple conditioning images. This raises a key question: *Can interleaved alignment be implemented directly on pretrained models without extensive unimodal pretraining?* Motivated by this question, we adopt a decoupled and scalable design that directly builds upon a pretrained MLLM and a diffusion transformer (DiT) pretrained on video generation. We name this framework DuetGen. DuetGen inherits the MLLM's visual understanding and world knowledge to generate text, while the video-pretrained DiT enables generation of image sequences with consistent objects and scenes. Concretely, the MLLM predicts a special token, <Begin-of-Vision> (BOV), to trigger image generation. To generate a new image, the previous images within the interleaved conversation history, either input or generated, are treated as conditioning frames for the DiT, while the MLLM hidden states preceding the <BOV> token provide semantic and linguistic guidance. This modular framework supports diverse choices of strong pretrained DiT and top-performing MLLMs without the need of unimodal pretraining from scratch and balancing understanding and generation objective in joint learning.

Together with the model design, we propose a two-stage decoupled training strategy that postpones interleaved pretraining while preserving the performance of the pretrained MLLM. In the first stage, we fine-tune only the MLLM using curated, high-quality interleaved generation data under next-token-prediction supervision. This stage teaches the MLLM to appropriately trigger image generation through <BOV> token and to continue text generation based on generated visuals. In the second stage, referred to as the interleaved context alignment stage, we freeze the MLLM parameters and update the DiT. Beyond the instruction-tuning data, this stage leverages large-scale interleaved alignment data, including interleaved image–text sequences that capture transitions between frames extracted from 5 million videos, as well as open-source image generation and editing samples.

We evaluate DuetGen on two public interleaved generation benchmarks: CoMM [10] and InterleavedBench [30], which cover diverse tasks such as how-to questions and story generation, as well as different input formats (*e.g.*, generation from scratch and continuation). In addition, we construct a new Interleaved Benchmark, focusing on diverse everyday problems. This benchmark leverages recent MLLMs capable of identifying fine-grained issues and includes the latest unified models such as NanoBanana [18] and Zebra-CoT [24] fine-tuned from Bagel [13]. Across all three benchmarks, DuetGen consistently outperforms previous open-source methods by a substantial margin across multiple metrics, including text quality, image fidelity, completeness, and image–context alignment. Moreover, DuetGen achieves significant gains on text-to-image and image-editing benchmarks compared to unified models like Bagel [13] and OmniGen2 [50], underscoring the benefits of leveraging pretrained MLLMs and video generation models. We will release both the model and dataset to facilitate future research on interleaved generation.

Our contribution can be summarized as follows:

- We curate a high-quality 298k instruction-tuning dataset for interleaved generation, along with large-scale interleaved-alignment data.
- We design a model architecture that leverages strong unimodal generation models and introduce a novel, decoupled training strategy.
- We propose a benchmark for evaluating interleaved generation and provide comprehensive comparisons with existing open-source and commercial models.

2. Related Work

Unified model. Unified models aim to support both text and image generation within one model. Starting from Chameleon [43], some works [12, 16, 28, 51] convert images into discrete tokens and unify language and text generation under next-token-prediction. Others, such as Transfusion [62], Bagel [13], and the Show-o series [53, 54], adopt a hybrid design that uses next-token prediction for text and diffusion for images. Another line of works use discrete-diffusion approaches to unify language and text generation, including MMaDA [57] and Lumina-DiMOO [55]. In terms of training strategy, early-fusion models [13, 43, 62] train from scratch on mixed text, images, and large-scale interleaved sequences, which requires substantial data and compute. In contrast, some works [9, 26, 37] fuse a pretrained MLLM with a pretrained generator via different connector designs. Given the high cost of early-fusion training, we follow the pretrained-fusion approach while noting that our data, evaluation, and training strategies are also applicable to early-fusion pipelines.

Interleaved generation model and datasets. Although unified models can generate both text and images, most still require users to specify the output modality and cannot seamlessly alternate between modalities to generate interleaved content. Early attempts [12, 16, 54] demonstrate simple story-telling and how-to cases without quantitatively benchmarking these capabilities, and their output resolution remains limited. CoMM [10] improves over noisy web-scale pretraining by converting how-to webpages into multimodal conversations. However, its data still contains stylistic noise (e.g., external links, inconsistent tone) and low-quality user-uploaded images, motivating the need for a more rigorous data pipeline. Visual chain-of-thought methods [19, 24, 41] further use generated images to assist reasoning, but their data focuses on several predefined tasks such as navigation or counting, limiting generalization ability. Based on these issues, we build a data engine that filters and rewrites web content using LLMs/MLLMs, and use high-quality synthetic data to improve visual fidelity and text-image alignment.

3. Interleaved Multimodal Training Data

The training data of DuetGen is divided into two parts: 1) high-quality interleaved multimodal conversations that teach models to follow user instructions; 2) interleaved image-text sequences for context alignment.

3.1. Instruction Tuning Data

High-quality instruction-tuning data for interleaved generation remains extremely limited. To overcome both the quality and diversity constraints of existing data, we construct an interleaved instruction dataset from two complementary sources that jointly cover realistic, embodied, and visually high-fidelity cases.

Data engine for websites. The data engine converts raw webpages into multimodal conversations. Similar to CoMM [10], we source data from public how-to and storytelling websites, but introduce extensive post-processing and filtering, as illustrated in Fig. 2. Specifically, we collect webpages from WikiHow [4], StoryBird [3], Instructables [2], and eHow [1], obtaining a total of 347k pages and retaining 268k after removing those containing only text or invalid images (e.g., QR codes, icons, advertisements). The main body of each webpage is converted into Markdown format for structured processing. Our pipeline consists of two major steps: (1) content rewriting and reorganization, and (2) conversion to user–assistant dialogue. First, we process text and images separately. Text passages are rewritten by an LLM to remove artifacts such as HTML tags, formatting errors, and external links. All images are captioned and categorized (e.g., natural photos, GUI screenshots, document pages), and invalid or irrelevant ones are discarded. To ensure coherence, we prompt an MLLM to remove duplicate or near-identical consecutive images and reorder image–text pairs so that each image appears after its corresponding description. Finally, a multimodal LLM transforms the cleaned image–text sequences into realistic instruction-style dialogues, where the user may optionally provide an image and the assistant responds step-by-step with interleaved reasoning and visual illustrations. In contrast to prior pipelines [10, 63] without further rewriting and reorganization, our data engine actively denoises, restructures, and dialogizes web content, producing clean interleaved data for instruction tuning.

Synthetic data. While the website-derived data provide feasible real-world solutions, their image quality, resolution, and step granularity vary widely—some are overly detailed, while others are too sparse—limiting the model's ability to generate high-resolution and visually appealing results. To address this, we employ the state-of-the-art commercial model Nano Banana [18] to synthesize high-quality interleaved data.

To enrich query diversity, we design a hierarchical query pool spanning eight broad everyday domains (e.g., Home & Living, Pets & Animal Caring). Domain annotators further refine these into 151 subcategories and compose about 10 seed questions per subcategory, yielding 1,500 seed prompts. Using OpenAI O3 [35] with the highest reasoning budget, we expand these into 15,270 diverse instructions. During the expansion, the base categrory and other subcategories are also provided to avoid duplication. The resulting prompts are fed into Nano Banana to generate corresponding image–text interleaved sequences. See supplementary for more details. In practice we find that Nano Banana performs particularly well on cooking-related tasks, we additionally sample 15k dish images from MM-Food-100k [14] as prompts for synthetic data generation.

In total, we obtain around 30k high-quality synthetic interleaved sequences, reserving 700 for evaluation. The website and synthetic data complement each other – the synthetic portion provides high-resolution, stylistically consistent, and aesthetically appealing visuals that facilitate stable model learning.

3.2. Interleaved Data For Context Alignment

The interleaved data used for context alignment focuses on teaching the model to generate images consistent with preceding images and text. Unlike instruction-tuning data, these samples do not require meaningful linguistic interactions between a user and an assistant, making them relatively easy to acquire at scale. We leverage two primary sources: video transition captions and various image-generation tasks. For video data, following Bagel [13], we collect 5 million raw videos and segment each into 5-second clips. All videos are pre-processed through scene detection and filtering to ensure temporal consistency within each segment. For every clip, we extract the first and last frame and annotate the transition using Qwen2.5-VL-32B [7], describing object motion, human actions, and camera movements. This converts raw videos into interleaved image–text sequences where the text explicitly explains the visual transition between frames. For image generation data, we aggregate open-source datasets including ShareGPT-4o-Image [8], NHR-Edit [21], OmniGen1&2 [50, 52], UniWorld-V1 [26], and Echo-4o [58], covering text-to-image, image editing, and multi-reference generation. Compared to video data, which typically captures smooth, subtle transitions, these datasets teach the model creative visual manipulation skills, such as adding, removing, or replacing objects and modifying backgrounds, which are also essential for general interleaved generation.

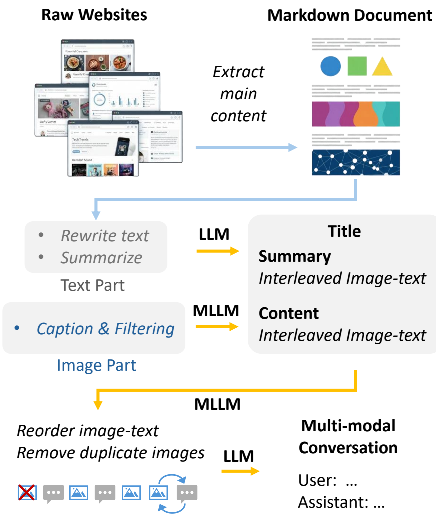

Figure 2. **Data engine for processing website data.** We design a data engine consists of a series of filtering and rewriting steps to convert noisy website data into high-quality instruction tuning data for interleaved generation.

4. Interleaved Generation Model

In this section, we introduce the architecture and training strategy of DuetGen. Prior interleaved generation models, such as Show-o2 [54] and Chameleon [43], adopt an early-fusion paradigm that jointly pretrains unimodal and interleaved generation abilities, requiring substantial effort to build both image understanding and image generation capabilities from scratch. In contrast, modern pretrained MLLMs and video generation models already provide strong multimodal reasoning and high-quality visual generation. This raises a natural question: can we directly leverage these pretrained capabilities and enable interleaved generation on top? To answer this, we design a framework that fuses a pretrained MLLM with a pretrained video generation model. Under this formulation, the unified model

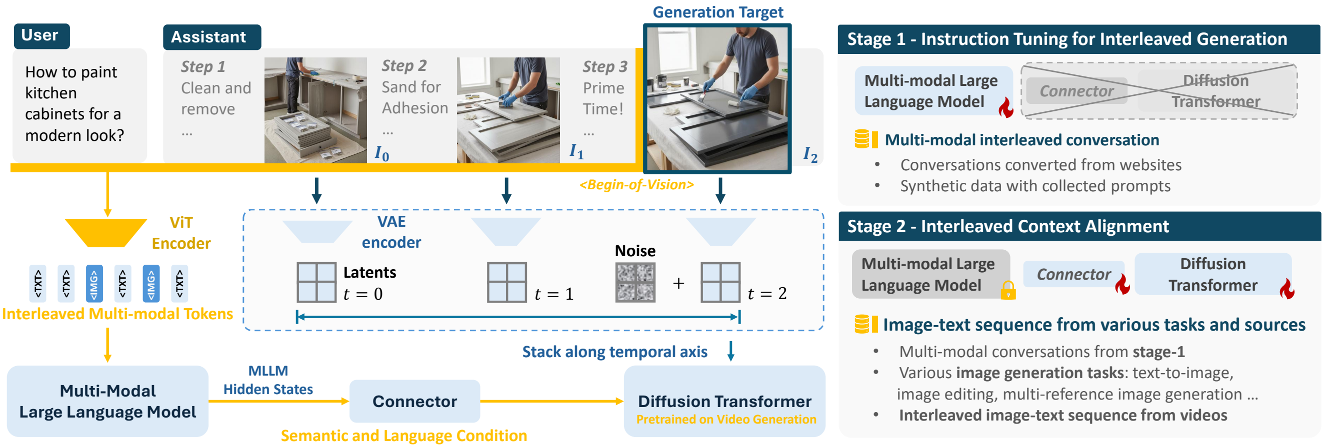

Figure 3. **Architecture and training strategy of DuetGen.** DuetGen consists of a pretrained multimodal large language model (MLLM) and diffusion transformer (DiT) pretrained on video generation. If a "<Begin-of-Vision>" (BOV) token is generated by the MLLM, then all the images in the interleaved sequence is packed as "condition frames" to the DiT and the MLLM hidden-states before the <BOV> token is sent to the DiT as the text condition to generate the new images.

only needs to learn two behaviors: (1) the MLLM must autonomously trigger image generation when visual predictions benefit reasoning or user tasks, and (2) the video generator must produce images consistent with prior text and images, whether user-provided or model-generated.

As shown in Fig. 3, DuetGen consists of an MLLM for text generation and a diffusion transformer (DiT) initialized from a video generation model for image synthesis. The MLLM can be any mainstream architecture equipped with a vision encoder and an LLM backbone, such as Qwen2.5-VL [7] or LLaVA [29]. The video generation component can be any model capable of conditioning on both images and text, such as Wan [44] or the Cosmos-Predict series [5].

During generation, the MLLM autoregressively predicts the next token. When a special `<BOV>` token (Begin-of-Vision) is generated, the model is switched into image-generation mode. Once `<BOV>` is produced, assume the preceding interleaved sequence is $T_1, I_1, T_2, I_2, \cdots, T_N$, consisting of both user-provided and previously generated multimodal content. Then the DiT part needs to generate image I_N conditioned on this sequence. For the visual latent input, we stack all images appearing before the `<BOV>` token along the temporal axis to form a set of conditioning frames, and encode them into latent embeddings using the VAE encoder. These latents are concatenated with the noisy latent of the target image to construct the visual input to the video generator. For the semantic and language condition, we extract the MLLM hidden states corresponding to all multimodal tokens preceding the `<BOV>` token. A lightweight connector projects these hidden states to the dimensionality required by the language-conditioning interface of the DiT.

During training, text generation is supervised with next-token prediction loss, masking out user input in the standard MLLM manner. The `<BOV>` token in the assistant turn is included in the loss, allowing the model to learn when to trigger image prediction. For image generation, we randomly sample one target image from each interleaved sequence, select a random diffusion step from the scheduler, and compute the loss (e.g., flow-matching [27]). During inference, the model autoregressively produces text until either a `<BOV>` token or the end-of-sequence token is reached. Once an image is generated, it is appended to the interleaved context, and the process repeats for subsequent steps. We further apply classifier-free guidance to enhance image fidelity: when generating the negative velocity, we keep the visual conditions fixed but remove the final text chunk from the MLLM hidden-state sequence.

4.1. Implementation Details

In this section, we introduce how to improve training efficiency. We adopt Qwen2.5-VL 7B [7] architecture for the MLLM and initialize the DiT backbone using Cosmos Predict 2.5 (2B) [6].

Packed sequence training. Sequence packing has become standard in MLLM/LLM training, allowing samples of different lengths and image resolutions to be packed together without padding and thereby improving training efficiency. However, the original implementation of Cosmos Predict 2.5 [40] is incompatible with interleaved samples containing images of heterogeneous sizes. To enable packed training, we introduce the following modifications: 1) For each interleaved sample, all images – regardless of resolution – are extracted and treated as a heterogeneous sequence of "video" frames. Their VAE latents are flattened and concatenated. For each image, we record its height, width, and index to restore the spatial shape during decoding; 2) We extend the original position embedding implementation. Now temporal indices increase by one after each image in the interleaved sequence, and the spatial RoPE (height/width indices) is computed using the per-image resolution.

Condition input. For text conditioning, guidance is injected via cross-attention between the image latents and the language embedding at every DiT decoder layer. Following Wang *et al.* [45], we concatenate the hidden states from all decoder layers along channel dimensiton to enhance representation. To prevent out-of-memory issues, we cap the maximum side length of images fed into the MLLM at 480 pixels. For visual conditioning, Cosmos Predict 2.5 concatenates the condition image latents with the noisy latents of the target frame along the temporal axis. We adopt the same strategy: the clean latents of user-provided images and previously generated images are concatenated with the noisy latent corresponding to the current generation target, forming a unified visual condition sequence.

4.2. Decoupled Training Strategy

Based on our interleaved data and model architecture, we adopt a decoupled two-stage training strategy. As illustrated in Fig. 3, training is divided into: (1) instruction tuning of the MLLM for interleaved generation, and (2) interleaved context alignment for the connector and DiT. In the first stage, we update only the MLLM parameters using the high-quality multimodal conversations described in Sec. 3.1, supervised with next-token prediction. After this stage, the MLLM learns to autonomously trigger image generation at appropriate moments and to continue text generation conditioned on newly produced images. We intentionally exclude data for interleaved context alignment discussed in Sec. 3.2 here, as such data lacks meaningful user–assistant interactions; introducing it too early may harm the pretrained MLLM's carefully engineered post-training behaviors. In the second stage, we freeze the MLLM and fine-tune only the connector and DiT, using the context-alignment data from Sec. 3.2, which includes video-labeled interleaved sequences and diverse image generation/editing datasets. We also add the instruction tuning data in Sec. 3.1 into the training. This enables image generation that stays well aligned with preceding images and textual context. This decoupled approach also let us leverage heterogeneous data effectively: even if text from the alignment data may be uninformative for a strong pretrained MLLM, it remains valuable for aligning visual generation behavior. The same strategy is applicable to other unified frameworks with separated language and diffusion parameters, such as Bagel [13].

5. Experiment

In this section, we present results on interleaved generation benchmarks, followed by evaluations on image generation and image editing tasks. We additionally conduct ablation studies on different data recipe to validate the effectiveness of our data engine and the contribution of synthetic interleaved data.

5.1. Interleaved Generation

We evaluate interleaved generation on our benchmark and two public benchmarks: CoMM [10] and Interleaved-Bench [30]. CoMM [10] contains text-only instructions covering story generation and how-to questions. InterleavedBench extends this setting with more tasks like passage generation and additional input formats like continuation tasks where models complete partially provided interleaved contexts. Both benchmarks rely on GPT-4o [33] for evaluation, scoring text and image completeness, image coherence, and image–text alignment. However, we observe that GPT-4o often misses fine-grained visual artifacts or subtle mismatches between user context and generated visuals.

To better evaluate results rigorously, we introduce a new benchmark, focusing on diverse, realistic tasks and employing stronger VLMs for evaluation. It includes two subsets: Cooking-200, where the model generates interleaved recipes consistent with a provided dish image and title, and How-to-500, an open-ended set of 500 everyday questions spanning 151 subcategories. Using an MLLM-as-judge protocol, we find that GPT-5 [34] reliably identifies subtle visual–semantic inconsistencies—for example, penalizing cases where the model generates headless shrimp despite the input image clearly showing shrimp with heads—while GPT-4o often overlooks such fine-grained errors. We report both sequence-level metrics (text completeness, image completeness, image coherence) and image-level metrics (aesthetic quality, image–text coherence). Refer to supplementary for more details.

Quantitative Comparison. Tables 2 and 3 present results on CoMM [10] and InterleavedBench [30]. DuetGen consistently outperforms prior systems across all major dimensions, including text quality, image quality, visual coherence, and image–text alignment. On the CoMM test set, it achieves a substantial improvement in Illustration Relevance Score (IRS) measuring image-text alignment, reaching 2.8× the score of the second-best method (7.76 vs. 2.71 for MiniGPT-5). A similar trend is observed on InterleavedBench with continuation tasks that require interpreting user-provided images and contextual inputs, where DuetGen shows an even larger advantage in text quality, attaining 3.4× the score of Emu2. These results show that DuetGen can comprehend complex user inputs to generate coherent and helpful textual solutions, and produce high-

Table 1. **Comparison on interleaved generation tasks.** T-Com, I-Com, I-Co, IT-Co, I-Q denotes text completeness, image completeness, image-coherence, image-text coherence, and image quality, respectively. 7B/2B in size column denotes the activated parameters for text generation and image generation if using decoupled design.

Model	Size	Cooking-200					Cooking-200-Text-Input					How-to-500				
		T-Com	I-Com	I-Co	I-Q	IT-co	T-Com	I-Com	I-Co	I-Q	IT-Co	T-Com	I-Com	I-Co	I-Q	IT-Co
Nano Banana [18]	-	4.24	4.07	4.36	4.81	4.83	4.02	4.02	4.59	4.75	4.72	3.95	4.28	4.49	4.22	4.24
SEED-LLaMA [16]	7B/0.8B	1.99	1.63	2.93	3.14	1.65	2.08	1.70	2.99	3.35	1.86	1.61	1.50	3.18	2.97	1.69
MiniGPT-5 [61]	7B/0.8B	1.85	1.81	1.75	2.81	1.88	2.03	2.27	1.84	2.91	2.10	1.94	2.22	2.63	2.98	2.43
Zebra-CoT [24]	7B/7B	2.10	2.63	3.54	3.61	3.67	2.12	2.54	3.21	3.06	3.07	2.04	2.05	3.52	2.84	2.59
DuetGen	7B/2B	**3.61**	**4.70**	**3.92**	**4.78**	**4.75**	**3.82**	**4.77**	**4.17**	**4.79**	**4.76**	**3.39**	**4.22**	**4.21**	**4.08**	**4.18**

Table 2. **Comparison on CoMM [10].** Sty. and Enti. denotes the style and entity consistency among generated images. Tren. denotes the trend alignment betwen image and text squence. Comp. denotes the completeness, ImgQ is the image quality. IRS is the illustration relevance score which is used to measure whether the generated images fits the surrounding context.

Model	Sty.	Enti.	Tren.	Comp.	ImgQ	IRS
MiniGPT-5 [61]	5.65	5.2	5.25	5.81	6.15	2.71
SEED-LLaMA [16]	7.55	6.81	6.15	5.13	6.36	1.46
Emu2 [42]	8.41	7.56	7.63	7.54	7.59	2.02
DuetGen	**9.22**	**9.22**	**9.24**	**9.66**	**9.53**	**7.76**

Table 3. **Comparison on InterleavedBench.** T-Q, I-Q, I-Co, IT-Co denotes text-quality, image-quality, image-coherence and the image-text coherence, respectively.

Model	T-Q	I-Q	I-Co	IT-Co	Helpfulness	Avg.
MiniGPT-5 [61]	1.22	2.45	1.62	2.03	1.77	1.82
GILL [20]	0.75	3.21	2.25	1.53	1.48	1.84
Emu2 [42]	1.26	2.28	1.89	1.34	1.64	1.68
DuetGen	**4.28**	**3.65**	**3.70**	**3.69**	**4.06**	**3.87**

quality images that remain closely aligned with the accompanying text, demonstrating the advantages of utilizing well-pretrained models.

Table 1 reports results on the two subsets of our benchmarks. Cooking-200-Text-Input removes images from user input. Nano Banana [18] shows strong performance across all the subsets, especially on How-to-500, which requires broader knowledge and the ability to generate physically plausible objects and procedures. DuetGen surpasses all other open-source models by large margins across all metrics, with particularly notable gains on How-to-500. Moreover, DuetGen significantly narrows the gap between open-source models and Nano Banana; on the more constrained Cooking-200 tasks, DuetGen even matches Nano Banana on certain metrics such as image–text coherence. These results highlight the potential of our framework: with sufficient high-quality data, DuetGen can approach the performance of top commercial models on specific domains.

Qualitative Results. Fig. 1 presents two interleaved generation examples. The model produces high-resolution images (768×768) with fine visual details and strong consistency both across generated frames and between user inputs and model outputs. In the grilled-steak example, DuetGen idenitify and generate the sides such as tomatoes, broccoli, and potatoes. In the bathe-a-cat example, the model maintains consistency of major objects—including the bathroom environment, human, and the cat—across multiple steps, demonstrating robust spatial and semantic coherence during interleaved reasoning and generation. Additional examples are provided in the supplementary material.

5.2. Image Generation and Editing

We use GenEval [17] to evaluate the image generation capabilities and use ImgEdit [59] and the English subset of GEdit [31] to evaluate the image editing performance.

Image Generation. Our method significantly outperforms other unified generation models on the overall score. In particular, DuetGen achieves strong improvements on multi-object metrics such as counting (0.94), position (0.84), and attribute binding (0.80)—areas where unified models typically struggle—indicating enhanced compositional reasoning and spatial grounding. Overall, our approach substantially narrows the gap with state-of-the-art commercial and task-specialized generative systems, while establishing a new performance baseline for unified multimodal generation. See supplementary for detailed results.

Image Editing. Table 5 shows the result on ImgEdit [59] and GEdit_EN [31] benchmarks, which uses VLM to evaluate editing results from prompt following and visual quality. On ImgEdit [59], DuetGen significantly outperforms prior unified models, especially on more complex tasks like “hybrid”, “add”, and “replace”. While recent editing model, Qwen-Image-Edit, still achieves the highest overall score, DuetGen is narrowing the gap as a interleaved generation model, and shows better score on Remove (4.71), Replace (4.69), and Add (4.53), demonstrating strong capability in precise object-level transformations. GEdit_EN benchmark

Table 4. **Comparison on GenEval.** * denotes LLM prompt rewriting.** uses interleaved generation to improve image quality.

Model Type	Method	Overall
Commercial	GPT-4o-Image [33]	0.84
Generation	SDXL [38]	0.55
	DALLE-3 [32]	0.67
	FLUX.1-dev [22]	0.82
	Qwen-Image [49]	0.87
Unified Model	Emu3 [47]	0.54
	Show-o [53]	0.53
	Janus-Pro-7B [11]	0.80
	MMaDA [57]	0.63
	MetaQuery-XL* [37]	0.80
	Blip-3o [9]	0.84
	Bagel [13]	0.82
	UniWorld-V1 [26]	0.80
	OmniGen2 [50]	0.80
Interleaved Generation	Uni-CoT** [39]	0.83
	DuetGen	0.88

Table 5. **Combined comparison on ImgEdit and GEdit_EN.** G_SC, G_PQ, and G_O are sub-metrics for GEdit_EN.

Model Type	Method	ImgEdit Overall	GEdit_EN G_SC	G_PQ	G_O
Commercial	Nano Banana [18]	4.23	7.28	7.83	6.93
	GPT-4o-Image [33]	4.20	7.85	7.62	7.53
Generation	ICEdit [60]	3.05	5.11	6.85	4.84
	Step1X-Edit [31]	3.06	7.09	6.76	6.701
	FLUX.1 Kontext [Pro] [23]	4.00	7.02	7.6	6.56
	Qwen-Image-Edit [49]	4.27	8.00	7.86	7.56
Unified Model	OmniGen [52]	2.96	5.96	5.89	5.06
	Bagel [13]	3.20	7.36	6.83	6.52
	UniWorld-V1 [26]	3.26	4.93	7.43	4.85
	OmniGen2 [50]	3.44	7.16	6.77	6.41
	OVIS-U1 [46]	4.00	-	-	6.42
Interleaved Generation	Uni-CoT [39]	-	7.91	6.24	6.74
	DuetGen	4.19	7.68	7.76	7.35

Table 6. **Comparison of different data strategies.** Abbreviation is aligned with Table 2.

Data Configuration	Sty.	Enti.	Tren.	Comp.	ImgQ.	IRS
CoMM original	6.14	6.21	6.52	6.45	6.30	4.42
w. Our data engine	7.85	7.76	7.22	8.15	7.79	5.91
+ Synthetic Data	9.15	9.21	9.30	9.45	9.48	7.58

uses two metrics, "G_SC" is for semantic consistency which evaluate whether editing is consistent with user prompt and "G_PQ" is pixel quality. "G_O" is the geometry average of "G_SC" and "G_PQ". Our model achieves strong performance across the three metrics compared with other unified model and closely matching on commercial models and strong editing models. The results on image generation and editing demonstrating the advantage of building upon DiT well-pretrained on video generation, which offers good pixel generation quality and content creation abilities.

Qualitative Examples. Fig. 1 showcases more complex cases beyond the primitive editing operations covered in the benchmark. DuetGen can execute intricate instructions that simultaneously modify backgrounds, adjust character appearance or clothing, change age or pose, and alter overall visual style. In addition, DuetGen supports combining multiple reference images with different resolutions, enabling flexible and compositional editing. Additional results are provided in the supplementary material.

5.3. Data Ablation

In this section, we evaluate the effectiveness of our data strategy using three configurations of instruction-tuning data on the CoMM benchmark [10]: (1) the original CoMM data (with 200k remaining samples due to expired image links); (2) CoMM data processed using our data engine; and (3) the processed CoMM data further augmented with our synthetic interleaved data. As shown in Table 6, applying our data engine yields substantial gains in both text quality and IRS (image-text alignment), highlighting the benefits of MLLM-based post-processing and cleaning. Incorporating synthetic data provides additional improvements, especially in image quality and temporal–semantic consistency.

6. Conclusion

We present DuetGen, a framework that advances interleaved multimodal generation through high-quality data, architecture design, training strategy, and quantitative benchmark. We curate 298k high-quality instruction-tuning samples from complementary sources for interleaved generation, along with large-scale interleaved context data for alignment. Instead of coupling uni-modal and interleaved generation during pretraining, DuetGen directly leverages a well-pretrained MLLM and a DiT pretrained on video generation, avoiding the uni-modal pretraining efforts. Finally, we introduce a dedicated benchmark for interleaved generation, enabling more comprehensive and standardized evaluation of this under-explored task.

Limitations. Our work primarily focuses on enabling and evaluating interleaved generation, leaving several directions for future exploration. These include scaling up the pretrained components for stronger reasoning and visual fidelity, as well as conducting deeper ablations on architectural choices—such as comparing cross-attention–based conditioning with MMDiT-style designs [15].

References

[1] ehow. https://www.ehow.com/, 2025. Online resource offering how-to guides and instructions. 3

[2] Instructables. [https://www.instructables.com/, 2025. Online community sharing DIY projects and tutorials. 3

[3] Storybird. https://storybird.com/, 2025. Online storytelling and creative writing platform. 3

[4] wikihow. https://www.wikihow.com/, 2025. Online how-to community and instructional website. 3

[5] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 5

[6] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025. 5

[7] 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. 4, 5

[8] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025. 4

[9] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 3, 8

[10] 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. In *CVPR*, pages 8073–8082, 2025. 3, 4, 6, 7, 8

[11] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 8

[12] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 3

[13] 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. 1, 2, 3, 4, 6, 8

[14] Yi Dong, Yusuke Muraoka, Scott Shi, and Yi Zhang. Mm-food-100k: A 100,000-sample multimodal food intelligence dataset with verifiable provenance. *arXiv preprint arXiv:2508.10429*, 2025. 4

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 8

[16] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 2, 3, 7

[17] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023. 7

[18] Google DeepMind. Gemini 2.5 flash image ("nano banana") — image generation & editing model, Aug. 2025. Released August 26, 2025; state-of-the-art image generation and editing model. 2, 3, 4, 7, 8

[19] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025. 2, 3

[20] Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. Generating images with multimodal language models. In *NeurIPS*, 2023. 2, 7

[21] Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. Nohumansrequired: Autonomous high-quality image editing triplet mining. *arXiv preprint arXiv:2507.14119*, 2025. 4

[22] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 8

[23] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 8

[24] Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025. 2, 3, 7

[25] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, Jiashuo Yu, Hao Tian, Jiasheng Zhou, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, and et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. In *ICLR*, 2025. 2

[26] 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. 2, 3, 4, 8

[27] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 5

[28] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 3

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *NeurIPS*, 2023. 2, 5

[30] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 22002–22016. Association for Computational Linguistics, 2024. 3, 6

[31] 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. 7, 8

[32] OpenAI. Dall·e 3. https://openai.com/blog/dall-e-3, 2023. 8

[33] OpenAI. Gpt-4o: The next-generation omni model, May 2024. Introduces GPT-4o, a multimodal model capable of real-time audio, vision, and language. 6, 8

[34] OpenAI. Gpt-5 announcement, Jan. 2025. Announcement of GPT-5 availability; advanced multimodal and reasoning capabilities. 6

[35] OpenAI. Openai o3 and o4-mini system card. System card, OpenAI, Apr. 2025. 2, 4

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 2

[37] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 2, 3, 8

[38] 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. 8

[39] Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025. 8

[40] NVIDIA Research. Cosmos-predict2: Deep imagination research. https://research.nvidia.com/labs/dir/cosmos-predict2/, 2023. Accessed: 2025-09-14. 5

[41] Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025. 2, 3

[42] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 7

[43] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 3, 4

[44] Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 5

[45] Andrew Z. Wang, Songwei Ge, Tero Karras, Ming-Yu Liu, and Yogesh Balaji. A comprehensive study of decoder-only llms for text-to-image generation. In *CVPR*, pages 28575–28585, 2025. 6

[46] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025. 8

[47] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 8

[48] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022. 2

[49] 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*, 2025. 8

[50] 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. 2, 3, 4, 8

[51] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: a unified foundation model integrating visual understanding and generation. In *ICLR*, 2025. 3

[52] 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 *CVPR*, pages 13294–13304, 2025. 4, 8

[53] 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. In *ICLR*, 2025. 2, 3, 8

[54] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 3, 4

[55] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al.

Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025. 3

[56] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025. 2

[57] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025. 3, 8

[58] 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. 4

[59] 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. 7

[60] 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. 8

[61] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 7

[62] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025. 3

[63] 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. In *NeurIPS*, 2023. 2, 4