SIGMA: Selective-Interleaved Generation with Multi-Attribute Tokens

Xiaoyan Zhang[1,2] Zechen Bai[4] Haofan Wang[3] Yiren Song[4†]
[1] Creatly AI [2] University of Michigan [3] Lovart AI [4] National University of Singapore

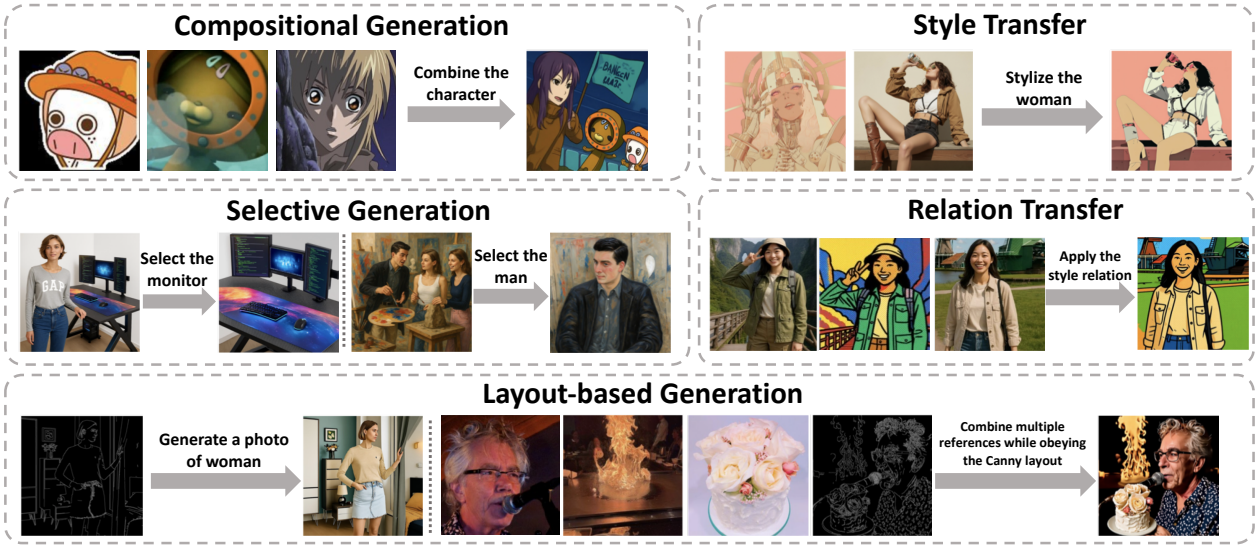


Figure 1. Overview of tasks within our unified framework, covering diverse generation scenarios including compositional generation, selective generation, stylization, style relation transfer, and layout-based generation.

Abstract

Recent unified models such as Bagel demonstrate that paired image–edit data can effectively align multiple visual tasks within a single diffusion transformer. However, these models remain limited to single-condition inputs and lack the flexibility needed to synthesize results from multiple heterogeneous sources. We present SIGMA (Selective-Interleaved Generation with Multi-Attribute Tokens), a unified post-training framework that enables interleaved multi-condition generation within diffusion transformers. SIGMA introduces selective multi-attribute tokens, including style, content, subject and identity tokens, which allow the model to interpret and compose multiple visual conditions in an interleaved text–image sequence. Through post-training on the Bagel unified backbone with 700K interleaved examples, SIGMA supports compositional editing, selective attribute transfer and fine-grained multimodal alignment. Extensive experiments show that SIGMA improves controllability, cross-condition consistency and visual quality across diverse editing and generation tasks, with substantial gains over Bagel on compositional tasks. Code is available at https://github.com/auihund/SIGMA.

† Corresponding author.

1. Introduction

Unified generative models [16, 20, 27, 53–55, 57] have recently emerged as a powerful paradigm for combining diverse visual tasks within a single architecture. The Bagel model [8] achieves impressive generalization in unified image generation and editing by training on paired image–edit data, allowing one diffusion transformer to perform both synthesis and manipulation. However, such unified models are typically restricted to single-condition inputs—for example, a single reference image or prompt—thus limiting their ability to compose information from multiple sources. In many practical scenarios, generation requires combining heterogeneous conditions, such as identity, content, and artistic style, into one coherent visual result.

This issue of binding [14], or how elements from different sources can be combined in a unified representation, has been a longstanding challenge in representation learning. In perceptual modeling, object-centric frameworks [28, 30, 58] address this problem by disentangling and tracking object-specific representations across scenes. However, in generative and editing settings, the binding problem reappears in a different form, where models must learn to as-

sociate semantic attributes such as identity, style, or layout with their visual targets. Earlier approaches [19, 25, 39] typically achieved binding through task-specific architectural designs (e.g., using separate encoders for content and style) or by overfitting to a single editing modality. These methods often fail to generalize across diverse conditions, especially in auto-regressive architectures like transformers, which are key to models like Bagel [8].

To address this limitation, we propose SIGMA (Selective-Interleaved Generation with Multi-Attribute Tokens), a post-training framework that extends the unified Bagel model to handle interleaved multi-condition generation. SIGMA introduces a mechanism for binding that allows users to upload multiple condition images, such as a person photo, an accessory image, and a style reference, and describe their relationships through interleaved text–image sequences (e.g., "a photo of a man" + his portrait + "with a photo of a dog" + the dog image + "in the style of Van Gogh" + a style image). The diffusion transformer then interprets this mixed sequence to produce a coherent, attribute-composed image, effectively bridging the gap between multi-reference input and unified generation.

At the core of SIGMA is the introduction of multi-attribute tokens, which enable the model to selectively control which aspects of each condition are used during generation. We design specialized tokens for attributes such as style, content, identity, and subject, allowing fine-grained control over how different references contribute to the final output. For instance, when a portrait of Van Gogh is encoded under a Style Token, the system extracts his artistic brushwork, while encoding it under an Identity Token transfers his facial identity. This selective conditioning mechanism empowers SIGMA to perform attribute-specific reasoning and cross-modal composition, enabling a form of binding that is essential for flexible image editing and synthesis which are not supported by existing unified models.

We further perform interleaved post-training on the Bagel backbone using a newly collected dataset of 700K interleaved examples, spanning diverse combinations of reference images, textual prompts, and style–content mappings. This post-training not only enhances compositional understanding but also allows flexible user interaction, supporting mixed text–image inputs and partial editing without fine-tuning. Extensive experiments across multiple generation and editing benchmarks show that SIGMA substantially improves controllability, visual coherence, and attribute alignment. In particular, SIGMA achieves clear improvements over the Bagel unified model across compositional, selective, and layout-based generation, while approaching the performance of GPT–4o and Nano-Banana on several challenging benchmarks. An overview of the tasks supported by our unified framework is illustrated in Figure 1.

Our core contributions are:

- We propose SIGMA, a unified post-training framework based on Bagel that supports interleaved multi-condition generation through a diffusion transformer architecture.
- We introduce a set of selective tokens that enable fine-grained control over how multiple condition images are composed and interleaved during generation.
- We construct a 700K interleaved image–text dataset and conduct extensive experiments, demonstrating SIGMA's superior controllability, compositionality, and visual fidelity across various editing and synthesis tasks.

2. Related Work

2.1. Image Generation

Recent advances in generative modeling have enhanced the fidelity and controllability of image synthesis. Early frameworks based on Generative Adversarial Networks (GANs) [13, 23] demonstrated capabilities in generating high-resolution and realistic imagery, yet suffered from training instability and limited diversity. The introduction of Denoising Diffusion Models (DDMs) [17, 44] fundamentally reshaped the field, replacing adversarial training with iterative denoising. These models capture richer multi-scale visual statistics and enable fine-grained control through noise scheduling and classifier-free guidance. More recently, Diffusion Transformers (DiTs) [40] unified diffusion dynamics with transformer-based architectures, demonstrating superior scalability and cross-domain generalization [9, 18, 26, 29, 50, 51, 56, 70]. Building upon these foundations, subsequent works such as Flux [24], PixArt-α [4], and OmniGen [59] explored cross-modal conditioning and large-scale pretraining to achieve both visual quality and semantic alignment.

2.2. Conditional Generation

Conditioned image generation aims to control the generative process via external signals such as subject [15, 22, 49, 67, 71], layout [33–36, 60, 65, 69], font [31, 47, 48], or style cues [12, 52, 68]. Text-to-image diffusion models, including Stable Diffusion [44], Imagen [45], and DALL·E 2 [42], demonstrated unprecedented semantic controllability by aligning latent representations with language embeddings. Beyond textual prompts, multimodal conditioning strategies have emerged to guide generation via spatial and structural constraints such as ControlNet [65], T2I-Adapter [37], and IP-Adapter [62]. These methods enable precise manipulation over pose, depth, or sketch inputs. However, existing pipelines often rely on independent modules or task-specific fine-tuning, limiting cross-condition generalization.

2.3. Unified Generative Models

While task-specific conditional models achieve high controllability, they typically lack cross-domain generaliza-

tion and parameter efficiency. The emerging trend is to build unified generative frameworks that jointly support multiple conditioning modalities within a single backbone. Representative efforts include OmniControl [60], Make-A-Scene [11], and PixArt-Σ [5], which extend diffusion transformers to handle multimodal prompts such as text, layout, and sketches in a unified latent space. Other approaches, such as UniDiffuser [1] and Muse [2], jointly learn bidirectional mappings between vision and language, supporting both generation and understanding within the same model [64]. These unified paradigms highlight a shift toward general-purpose generative intelligence—models capable of seamlessly adapting to diverse conditions and tasks without task-specific retraining. Our work follows this trend by designing a unified diffusion–transformer architecture that maintains modality-consistent representations while preserving fine-grained conditional controllability.

3. Method

In this section, we describe the unified backbone and post-training setup in Sec. 3.1,the multi-attribute token design in Sec. 3.2, the interleaved conditioning mechanism in Sec. 3.3, and the group-scoped attention mask in Sec. 3.4. The overall architecture is illustrated in Fig. 2.

3.1. Unified Backbone and Post-Training

We build upon the unified model paradigm established by Bagel, which integrates multiple image generation and editing tasks within a single diffusion transformer backbone. Let $\mathcal{M}_\theta$ denote the pretrained unified diffusion model parameterized by θ, trained on paired data $(x_{\text{src}}, x_{\text{tgt}})$ representing pre- and post-edit images. Given an input image x_{src} and a text instruction c, the model predicts the target image x_{tgt} through the denoising process:

$$\mathbf{z}_{t-1} = \mathcal{M}_\theta(\mathbf{z}_t, c, x_{\text{src}}, t), \tag{1}$$

where $\mathbf{z}_t$ denotes the latent at timestep t. This unified setup allows multi-task learning across generation, editing, and inpainting.

However, the original unified model is restricted to single-condition inputs and cannot handle multimodal conditioning (e.g., multiple reference images with different semantics). To overcome this limitation, we introduce a **post-training phase** that augments the Bagel backbone to support *interleaved multi-condition inputs*. Instead of conditioning on a single pair (x_{src}, c), SIGMA is trained on interleaved sequences s that mix multiple reference images and text spans with multi-attribute tokens. We keep the same denoising objective as in the backbone and simply change the conditioning form:

$$\mathcal{L}_{\text{SIGMA}} = \mathbb{E}_{(s,x_{\text{tgt}}),\,t}\left[\left\|\mathbf{z}_{t-1} - \mathcal{M}_\theta(\mathbf{z}_t, s, t)\right\|_2^2\right], \tag{2}$$

where s denotes the interleaved text–image sequence that includes the text prompt, all condition images, and their associated attribute tokens.

3.2. Multi-Attribute Token Design

The key innovation of SIGMA lies in its **Selective Multi-Attribute Tokenization**. Instead of treating all conditioning images equally, we assign a specific *attribute token* to each condition according to its semantic role. Formally, for an input image x_i, we define a task-specific embedding $\tau_i \in \mathcal{T}$ chosen from a fixed attribute vocabulary:

$$\mathcal{T} = \{\texttt{Style}, \texttt{Subject}, \texttt{Identity}, \texttt{Layout}, \ldots\}. \tag{3}$$

Each token τ_i modulates the feature extraction and fusion process by controlling which latent subspace of the diffusion transformer is activated. This design allows selective extraction of visual attributes: for instance, encoding a Van Gogh portrait under a `Style` token captures brushstroke patterns, while encoding it under an `Identity` token preserves facial characteristics.

Given the encoded condition image features $\mathbf{v}_i = E_\phi(x_i)$ and the corresponding attribute token τ_i, we compute a token-conditioned embedding:

$$\mathbf{t}_i = \mathbf{v}_i + W_\tau(\tau_i), \tag{4}$$

where W_τ is a learnable attribute projection matrix. These attribute-specific tokens provide fine-grained control and facilitate multi-attribute composition during generation.

3.3. Interleaved Conditioning Mechanism

To support multi-condition fusion, we introduce an **interleaved conditioning** mechanism that allows text and image inputs to appear in an alternating sequence. Let $\mathbf{T}_k$ denote text embeddings and $\mathbf{I}_k$ denote the corresponding image condition embeddings with assigned attribute tokens. The final input sequence to the diffusion transformer is constructed as:

$$\mathbf{H} = [\mathbf{T}_1; \mathbf{I}_1; \mathbf{T}_2; \mathbf{I}_2; \ldots; \mathbf{T}_n; \mathbf{I}_n], \tag{5}$$

where $[\cdot]$ denotes token concatenation in the temporal order of user specification. This interleaved structure enables flexible multimodal reasoning: the model can learn to parse textual descriptions and visual conditions jointly, preserving contextual alignment across modalities.

During training, the diffusion transformer processes this interleaved sequence through a series of cross-attention and self-attention layers. The Group-Scoped Attention Mask (Section 3.4) ensures that only relevant attribute tokens influence specific attention heads, while implicit alignment feedback encourages consistent blending between adjacent conditions. This alignment signal naturally arises from the denoising objective and the interleaved token structure, guiding the model to associate each attribute token with its corresponding visual source without requiring an explicit reward model.

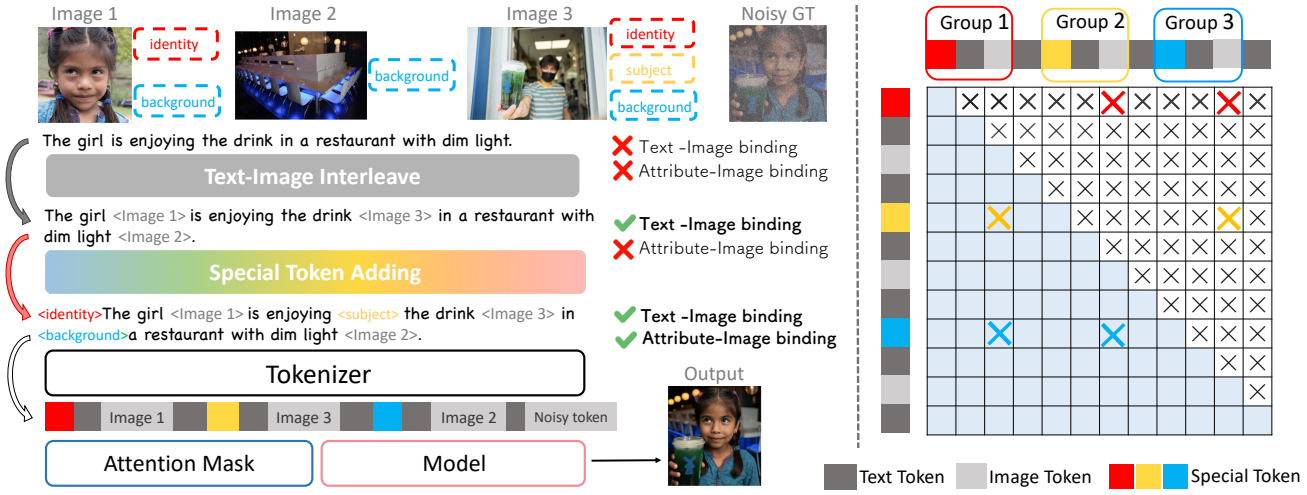


Figure 2. Overview of SIGMA. **Left:** Each sample may include multiple images, and each image can involve multiple visual attributes. Text–Image Interleave aligns textual spans with image placeholders, while Special Token Adding binds specific attributes to their corresponding images, avoiding semantic entanglement. **Right:** On top of causal attention, we add group-scoped masks so each special token can only attend to images within its own group, effectively reducing cross-image interference and ensuring clean attribute–image binding.

At inference, users can freely combine different types of conditions (e.g., "a dog photo + a style painting + an environment image") in any interleaved order. SIGMA interprets each condition adaptively based on its token semantics, producing coherent and controllable compositions across styles, identities, and layouts.

3.4. Group-Scoped Attention Mask

Although interleaved conditioning enables flexible multimodal composition, it also introduces undesired *attribute leakage*, where special tokens corresponding to one reference image may attend to unrelated image patches from other conditions. Such uncontrolled attention often leads to semantic confusion between reference signals, resulting in degraded controllability and inconsistent generation. To address this, we design a simple but effective **group-scoped attention mask** that restricts cross-group attention for special tokens while preserving other attention patterns, thereby retaining the model's ability to reason about spatial layouts, geometry, and alignment across conditions.

We denote the final input sequence by $\mathbf{H} = \{h_1, \ldots, h_L\}$, where L is the total number of tokens. Each token h_ℓ has a type $\mathrm{type}(h_\ell) \in \{\texttt{special}, \texttt{text}, \texttt{image}, \texttt{plain}\}$, representing attribute tokens (e.g., <id>, <style>), textual instruction tokens, image patch tokens, and other non-grouped text tokens, respectively. Tokens of type `special` or `image` are additionally associated with a group index $\mathrm{grp}(h_\ell) \in \{1, \ldots, m\}$, corresponding to each reference condition. We construct a binary attention mask $\mathbf{B} \in \{0,1\}^{L \times L}$, where each element $\mathbf{B}[q,k]$ determines whether the query token h_q is allowed to attend to the key token h_k. The final mask is obtained by combining three components:

$$\mathbf{B} = (\mathbf{C} \wedge \mathbf{M}) \vee \mathbf{S}, \qquad \mathbf{A} = (\mathbf{1} - \mathbf{B}) \cdot (-\infty), \tag{6}$$

where $\mathbf{A}$ is added to the attention logits before the softmax. Here $\mathbf{C}$ encodes the causal attention pattern inherited from Bagel, $\mathbf{S}$ enables unrestricted intra-image attention, and $\mathbf{M}$ enforces the group constraint.

The causal mask $\mathbf{C}$ maintains the autoregressive property of the original model by allowing each token to attend only to previous tokens in the sequence:

$$\mathbf{C}[q,k] = 1 \quad \text{iff} \quad k \le q. \tag{7}$$

The intra-image mask $\mathbf{S}$ enables full bidirectional attention among patch tokens belonging to the same image, which is crucial for local reasoning about structure, geometry, and spatial relations:

$$\mathbf{S}[q,k] = \begin{cases} 1, & \text{if } \mathrm{type}(h_q) = \mathrm{type}(h_k) = \texttt{image}, \\ & \quad \mathrm{img}(h_q) = \mathrm{img}(h_k), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Finally, the group constraint $\mathbf{M}$ restricts special tokens from attending to image patches outside their corresponding group, effectively blocking cross-condition leakage:

$$\mathbf{M}[q,k] = \begin{cases} 0, & \mathrm{type}(h_q) = \texttt{special}, \\ & \mathrm{type}(h_k) = \texttt{image}, \\ & \mathrm{grp}(h_q) \neq \mathrm{grp}(h_k), \\ 1, & \text{otherwise.} \end{cases} \tag{9}$$

This masking strategy enforces minimal but effective structure in the attention map. Special tokens are linked

only to their designated image patches. Images preserve unrestricted internal connectivity, and the rest of the sequence follows the original causal ordering. The design prevents attention drift across unrelated conditions while still allowing the model to capture global dependencies and relational cues through text tokens and causal connections. In practice, this yields attribute disentanglement and more controllable multi-condition generation without reducing the expressive capacity of the underlying diffusion transformer.

4. Interleaved Multi-Condition Dataset

4.1. Dataset Overview

To enable SIGMA to learn which attribute should be taken from which reference, we curate a large-scale interleaved dataset where text spans, special tokens, and multiple condition images are aligned. The final corpus contains **700K** sequences across six major task families—compositional generation (100K), selective content extraction (226K), stylization (153K), relation transfer (41.6K), image editing (70K), and conditional layout generation (110K). The distribution of these task families is illustrated in Fig. 3(a). Compared with standard caption–image pairs, this interleaved formulation forces the model to handle multi-image, multi-attribute, and asymmetric-reference scenarios, which are central to controllable multimodal generation.

To make attribute–image bindings machine-readable, we introduce 14 special tokens denoting entity- or attribute-level cues such as identity, subject, clothing, style, layout, pose, and lighting. Tokens are injected directly before the corresponding textual mention, followed by the referenced image. This converts each caption into a structured multimodal sequence that specifies which visual factor should be extracted from each source. Many samples are intentionally attribute-dense (e.g., `<subject>` + `<clothing>` + `<background>` for a single image), allowing SIGMA to learn disentangled selection rather than uncontrolled fusion when references overlap in content or appearance.

4.2. Dataset Construction

Our dataset combines newly generated samples and adapted high-quality corpora. We synthesize large-scale compositional, stylization, relational style-transfer, and editing data using GPT-4o and Nano-Banana, covering human, object, and scene combinations. The selective extraction subset is derived from Echo-4o by treating each compositional output as input and identifying extraction targets using GPT-4o. The conditional-layout subset incorporates canny/depth maps and layout-designated reference images, with geometric cues extracted by MiDaS. Existing datasets such as Nano-150K, Echo-4o, X2Edit, and ShareGPT-4o are converted into the interleaved format via token injection.

To unify data from heterogeneous sources, we employ a structured token-injection pipeline that converts each sample into an interleaved text–image sequence, as shown in Fig. 3(b). Each entity-bearing phrase in the caption is paired with a special token and its corresponding reference image, yielding locally bound text–image groups that specify the intended visual attributes and their sources. This standardized interleaving process makes cross-condition relationships explicit and enables SIGMA to learn fine-grained, source-aware attribute control.

5. Experiment

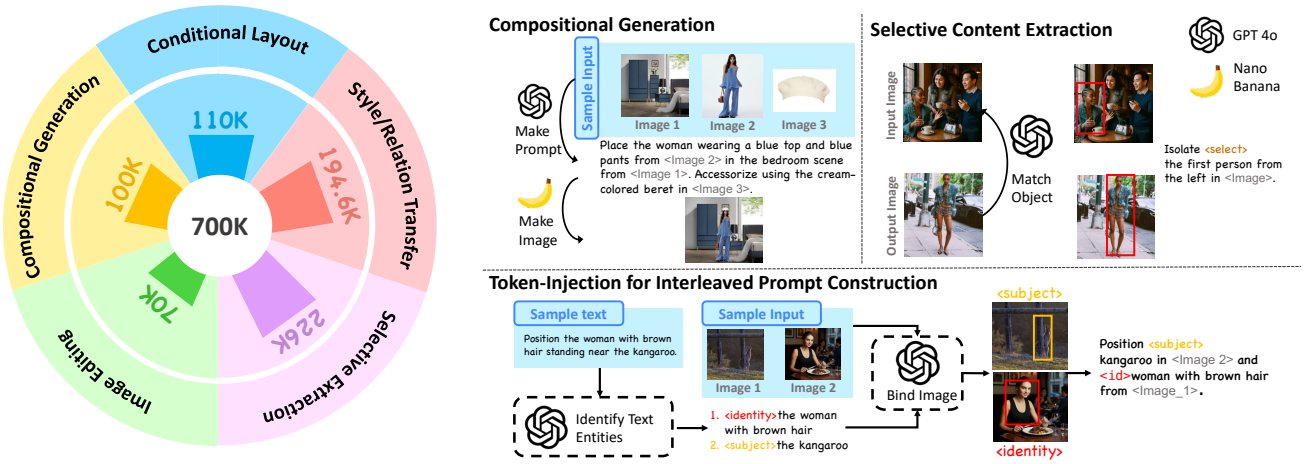
5.1. Experimental Setup

Training Details. We build our model upon the Bagel unified diffusion backbone [8], training only the generation branch while freezing the VAE module. For each task family in our interleaved multi-condition corpus, we allocate 95% of the sequences to the training set, where multiple *special tokens* serve as conditioning signals. The model is trained for 50K steps on 4 NVIDIA H200 GPUs. To improve efficiency, we apply token packing with a maximum of 30K tokens per packed batch. A cosine learning rate schedule is used with a peak learning rate of 2×10^{-5} and a minimum learning rate of 10^{-7}. Optimization is performed using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) with gradient clipping of 1.0, and we employ fully sharded data parallelism to scale training across multiple GPUs.

Benchmarks. We evaluate our method on several benchmarks that cover different aspects of controllable image generation. We use XVerseBench [3] for compositional generation, which provides diverse human identities, objects, animals references for assessing multi-entity reasoning. In addition, we construct a new **comprehensive benchmark** that jointly evaluates controllability, compositional reasoning and structural alignment across four representative tasks: compositional generation, selective generation, style transfer and layout-based generation. All samples are drawn from the held-out portion of our corpus to ensure a strict separation between training and evaluation. Full statistics, benchmark construction details and per-task configurations are provided in the supplementary material.

Baselines. In addition to **Bagel** [8], we include recent unified diffusion transformers such as **XVerse** [3] and **SSR** [66]. We also incorporate **EasyControl** for the layout based generation setting. To provide a reference to general-purpose multimodal systems, we report results from closed-source models **GPT-4o** [21] and **Nano-Banana** [7].

Metrics. We evaluate controllability and perceptual quality using a mix of semantic and visual metrics. **CLIP↑**, **CLIP-I↑** [41], and **DINO↑** [38] similarities measure



(a) Dataset Overview **(b) Dataset Construction Pipeline**

Figure 3. Overview and construction of the interleaved multi-condition dataset. **(a)** The 700K corpus spans six task families, including compositional generation, selective content extraction, stylization, relation transfer, image editing, and conditional layout. **(b)** Data are built via compositional generation (GPT-4o + Nano-Banana), selective content extraction (object matching with GPT-4o), and token-injection that binds text entities with reference images for interleaved supervision.

Benchmark	Method	CLIP↑	CLIP-I↑	DINO↑	DreamSim↑
XVerse Bench	GPT-4o	32.94	**77.74**	**66.42**	68.11
	Nano-Banana	32.45	75.50	65.10	**69.54**
	SSR	27.72	69.19	46.98	63.24
	XVerse	**33.94**	67.53	45.24	66.25
	Bagel	24.32	66.32	56.13	53.31
	SIGMA (Ours)	31.96 (+7.64)	75.57 (+9.25)	59.52 (+3.39)	67.87 (+14.56)
Our Bench	GPT-4o	**31.07**	77.93	63.58	**67.49**
	Nano-Banana	30.65	75.63	62.22	62.78
	SSR	28.61	65.34	54.11	52.65
	XVerse	32.33	44.15	42.76	54.63
	Bagel	17.91	52.52	41.62	43.27
	SIGMA (Ours)	30.29 (+12.38)	**78.94** (+26.42)	**64.08** (+22.46)	62.45 (+19.18)

Table 1. Comparison of compositional generation performance across two benchmarks. Numbers in parentheses show improvement over Bagel. Rows highlighted in gray denote closed-source models.

Method	CLIP↑	CLIP-I↑	CLIP-ES↓	AES↑
GPT–4o	25.84	80.14	60.22	5.882
Nano–Banana	25.48	77.88	62.63	5.804
SSR	25.46	71.68	58.78	**5.914**
XVerse	25.15	70.61	61.16	5.181
Bagel	23.49	70.61	67.90	5.209
SIGMA (Ours)	**25.90** (+2.41)	**80.26** (+9.65)	**58.02** (–9.88)	5.849 (+0.64)

Table 2. Selective generation results on our benchmark. Numbers in parentheses show improvement over Bagel.

text–image and subject alignment, while **CLIP-ES**↓ assesses subject exclusivity in multi-reference settings. **AES**↑ [46] evaluates the overall aesthetic appeal, and **DreamSim**↑ [10] quantifies perceptual similarity aligned with human visual judgment. For layout-based generation tasks, we additionally report the **F1**↑ score computed between the extracted and input edge maps in edge-conditioned generation to assess structural consistency, and the **FID**↓ to measure distributional fidelity.

Condition	Method	F1↑	FID↓	CLIP↑	AES↑
Layout only	GPT–4o	0.09	196.02	27.33	5.545
	Nano–Banana	0.12	189.46	26.50	**5.743**
	EasyControl	0.16	135.27	27.14	5.357
	Bagel	0.10	**103.72**	25.82	4.13
	SIGMA (Ours)	**0.44** (+0.34)	121.08 (+17.36)	26.35 (+0.53)	5.649 (+1.52)
Layout + Reference	GPT–4o	0.04	182.48	24.39	**6.01**
	Nano–Banana	0.04	161.91	24.86	5.912
	Bagel	0.25	188.83	24.22	4.54
	SIGMA (Ours)	**0.35** (+0.10)	**108.05** (–80.78)	**24.53** (+0.31)	5.774 (+1.23)

Table 3. Layout-based generation results on the layout-based subset of our benchmark. Numbers in parentheses show improvement over Bagel.

5.2. Quantitative Evaluation

We assess SIGMA on three representative tasks. Across all settings, SIGMA delivers strong and reliable performance.

On the compositional generation benchmark, Table 1 shows that SIGMA outperforms both SSR and Bagel across all four metrics. Relative to Bagel, SIGMA achieves substantial gains on CLIP, CLIP-I, DINO, and DreamSim, highlighting the effectiveness of our fusion mechanism in preserving identity and compositional integrity. Compared with XVerse, SIGMA attains slightly lower CLIP scores but consistently higher CLIP-I, DINO, and DreamSim, indicating stronger structural and perceptual alignment with the references. Among all open-source methods, SIGMA delivers the best overall performance and approaches the quality of GPT–4o and Nano-Banana, suggesting that it can accurately capture the intended objects and attributes while maintaining geometric fidelity and fine-grained visual detail. For selective generation, Table 2 highlights SIGMA's ability to pick out the correct target while maintaining visually pleasing results. The method obtains the lowest CLIP-

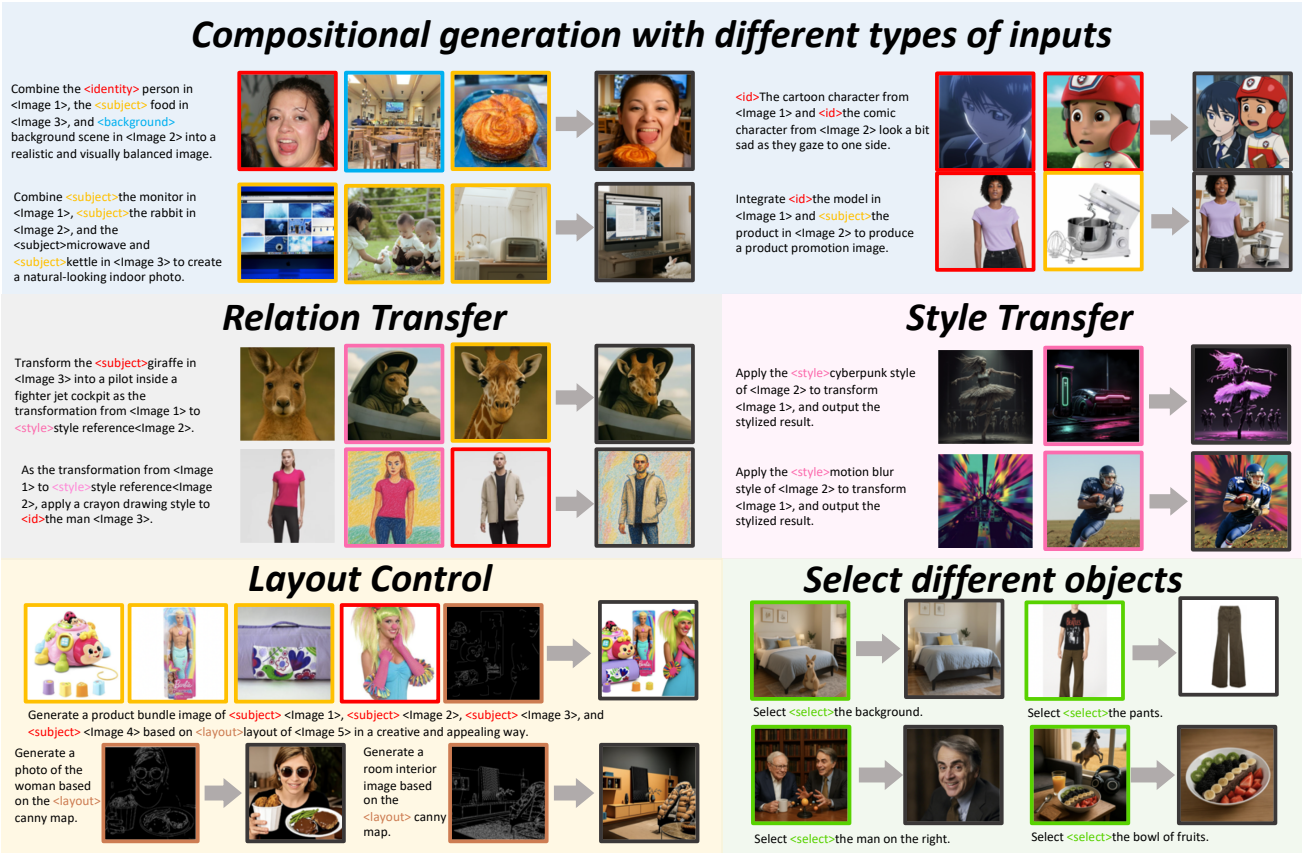


Figure 4. Qualitative results achieved by SIGMA. By leveraging specialized attribute tokens, SIGMA flexibly binds the required elements from input references, accomplishing a wide range of generation tasks. The results demonstrate clear binding between inputs and outputs, as well as high-quality, visually coherent generations across all scenarios.

ES score among all baselines, meaning it is less likely to select irrelevant objects. Together with its competitive CLIP and CLIP-I values, these results show that SIGMA can selectively extract and transform the desired object with dependable precision. For layout based generation, Table 3 shows that SIGMA follows layout constraints more accurately than prior unified models. The layout plus reference setting further demonstrates that SIGMA can align spatial structure with appearance guidance in a coherent manner.

5.3. Qualitative Evaluation

Figure 4 presents an overview of the visual capabilities enabled by SIGMA. The model produces high-quality results across a series of tasks and remains stable even in challenging multi-input scenarios. SIGMA can merge heterogeneous references, follow complex textual descriptions, and generate images that preserve fine-grained details such as texture, lighting, and identity. The outputs are visually coherent and maintain consistent composition across difficult cases, demonstrating that our multi-condition design provides reliable and flexible control over how different reference elements are incorporated into the final generation.

To further assess practical behavior, Figure 5 compares SIGMA with existing unified diffusion transformers. The examples illustrate performance under two demanding settings: compositional generation and selective generation. In the compositional setting, SIGMA can identify the correct objects from multiple visual references, integrate them naturally, and maintain stable geometrical relations and plausible scene structure. The generated images remain faithful to both the content and the text instructions. In comparison, SSR, XVerse, and Bagel often exhibit consistency issues, such as difficulty selecting the correct attributes and applying them properly during generation, while Nano maintains stronger consistency but sometimes produces unnatural shapes. The selective generation examples highlight another strength of the model. Each input image contains multiple candidate objects, yet SIGMA reliably follows the user instruction and extracts the intended target, whether it is a specific object or a background. The selected elements

Figure 5. Qualitative comparisons on our benchmark.

appear natural and well-formed, preserving identity and appearance without artifacts. At the same time, irrelevant objects are effectively excluded, indicating strong exclusivity and consistent instruction following.

Overall, the qualitative results show that SIGMA handles diverse tasks in a unified framework while producing coherent, detailed, and semantically aligned generations.

5.4. Ablation Study

Special	Mask	All	CLIP↑	CLIP-I↑	DreamSim↑	AES↑
		✓	25.85	62.67	44.74	5.576
✓		✓	29.25	74.26	57.11	5.561
✓	✓		27.32	72.64	58.65	5.506
✓	✓	✓	**30.29**	**78.94**	**62.45**	**5.731**

Table 4. Ablation of SIGMA components. "All" indicates full-parameter finetuning, while rows without a check in the "All" column correspond to LoRA-based tuning. We analyze the contributions of special tokens and group-scoped attention masks under both setups. Higher values indicate better alignment or perceptual quality.

We evaluate the contribution of multi-attribute tokens, the group-scoped attention mask, and the choice between full-parameter finetuning and LoRA tuning. As shown in Table 4, removing attribute tokens leads to a substantial drop in CLIP and CLIP-I, indicating that the model can no longer reliably determine which attributes should be extracted or where they should be applied. Adding the group-scoped mask yields further gains in CLIP-I and DreamSim, reflecting improved structural and perceptual consistency. LoRA-based tuning remains functional but trails full finetuning in both alignment and aesthetic quality, suggesting that limited parameter updates reduce the model's expressive capacity.

Figure 6. Qualitative ablation on attribute tokens, masked attention, and full-parameter training strategy.

Figure 6 provides the qualitative results. In the compositional generation example, omitting attribute tokens causes the model to confuse object roles entirely, while adding them without masking produces clearer objects but still unnatural interactions. LoRA tuning generates reasonable layouts but misses fine details such as missing cables. In the stylization case, the absence of the mask causes the model to collapse toward copying the style reference, showing that it fails to bind the style token to the content image. When the mask is included, attribute binding becomes correct and style transfer is coherent. LoRA again produces correct but less consistent stylization, whereas SIGMA achieves the highest fidelity across all elements.

6. Conclusion

We presented SIGMA, a unified diffusion transformer for controllable multi-condition image generation. The framework combines selective multi-attribute tokens that explicitly specify what to extract from each reference image, an interleaved text and image conditioning scheme that enables flexible multi-reference fusion, and a group-scoped attention mask that reduces unwanted interactions across different sources. These components are trained on a 700K interleaved dataset covering compositional generation, selective content extraction, stylization, relation transfer and layout-guided synthesis. This large and diverse corpus helps the model learn fine-grained attribute disentanglement and reliable reasoning across multiple inputs. Experiments on compositional, selective and layout-based generation demonstrate consistent improvements over unified baselines such as Bagel in terms of subject fidelity, structural alignment and controllability. Ablation results further verify that each part of the framework contributes to the overall performance. SIGMA offers a model-agnostic design that can be applied to future diffusion transformers and provides a practical foundation for the study of structured conditioning and unified multi-attribute generative modeling.

Acknowledgement

This work was supported in part by Creatly AI for providing resources and research support.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023. 3

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3

[3] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025. 5, 2, 3

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 3

[6] 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. 2

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5, 1, 3

[8] 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, 5, 3

[9] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 2

[10] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 6

[11] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European conference on computer vision*, pages 89–106. Springer, 2022. 3

[12] Yan Gong, Yiren Song, Yicheng Li, Chenglin Li, and Yin Zhang. Relationadapter: Learning and transferring visual relation with diffusion transformers. *arXiv preprint arXiv:2506.02528*, 2025. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[14] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. 1

[15] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19085–19096, 2025. 2

[16] Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*, 2025. 1

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[18] Shijie Huang, Yiren Song, Yuxuan Zhang, Hailong Guo, Xueyin Wang, and Jiaming Liu. Arteditor: Learning customized instructional image editor from few-shot examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17651–17662, 2025. 2

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[20] Zhipeng Huang, Shaobin Zhuang, Canmiao Fu, Binxin Yang, Ying Zhang, Chong Sun, Zhizheng Zhang, Yali Wang, Chen Li, and Zheng-Jun Zha. Wegen: A unified model for interactive multimodal generation as we chat. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23679–23689, 2025. 1

[21] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5, 1, 3

[22] Yuxin Jiang, Yuchao Gu, Yiren Song, Ivor Tsang, and Mike Zheng Shou. Personalized vision via visual in-context learning. *arXiv preprint arXiv:2509.25172*, 2025. 2

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[24] 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. 2

[25] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[26] Yuanhang Li, Yiren Song, Junzhe Bai, Xinran Liang, Hu Yang, Libiao Jin, and Qi Mao. Ic-effect: Precise and efficient video effects editing via in-context learning. *arXiv preprint arXiv:2512.15635*, 2025. 2

[27] Zongxia Li, Hongyang Du, Chengsong Huang, Xiyang Wu, Lantao Yu, Yicheng He, Jing Xie, Xiaomin Wu, Zhichao Liu, Jiarui Zhang, et al. Mm-zero: Self-evolving multi-model vision language models from zero data. *arXiv preprint arXiv:2603.09206*, 2026. 1

[28] Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407*, 2020. 1

[29] Cheng Liu, Yiren Song, Haofan Wang, and Mike Zheng Shou. Omnipsd: Layered psd generation with diffusion transformer. *arXiv preprint arXiv:2512.09247*, 2025. 2

[30] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020. 1

[31] Runnan Lu, Yuxuan Zhang, Jiaming Liu, Haofan Wang, and Yiren Song. Easytext: Controllable diffusion transformer for multilingual text rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7565–7573, 2026. 2

[32] Jian Ma, Xujie Zhu, Zihao Pan, Qirong Peng, Xu Guo, Chen Chen, and Haonan Lu. X2edit: Revisiting arbitrary-instruction image editing through self-constructed data and task-aware representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7764–7772, 2026. 2

[33] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2

[34] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

[35] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025.

[36] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 2

[37] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[40] 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

[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. 5

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 2

[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] 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

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 6

[47] Wenda Shi, Yiren Song, Zihan Rao, Dengming Zhang, Jiaming Liu, and Xingxing Zou. Wordcon: Word-level typography control in scene text rendering. *arXiv preprint arXiv:2506.21276*, 2025. 2

[48] Wenda Shi, Yiren Song, Dengming Zhang, Jiaming Liu, and Xingxing Zou. Fonts: Text rendering with typography and style controls. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18463–18474, 2025. 2

[49] Yiren Song, Shijie Huang, Chen Yao, Xiaojun Ye, Hai Ci, Jiaming Liu, Yuxuan Zhang, and Mike Zheng Shou. Processpainter: Learn painting process from sequence data. *arXiv preprint arXiv:2406.06062*, 2024. 2

[50] Yiren Song, Danze Chen, and Mike Zheng Shou. Layertracer: Cognitive-aligned layered svg synthesis via diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19731–19741, 2025. 2

[51] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572*, 2025. 2

[52] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025. 2

[53] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 1

[54] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[55] Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17001–17012, 2025. 1

[56] Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xiangtai Li, and Yiren Song. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv:2505.21541*, 2025. 2

[57] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025. 1

[58] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022. 1

[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

[60] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023. 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. 3

[62] 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. 2

[63] 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. 2

[64] Mingcheng Ye, Jiaming Liu, and Yiren Song. Loom: Diffusion-transformer for interleaved generation. *arXiv preprint arXiv:2512.18254*, 2025. 3

[65] 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. 2

[66] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024. 5, 3

[67] Yu Zhang, Jingyi Liu, Yiwei Shi, Qi Zhang, Duoqian Miao, Changwei Wang, and Longbing Cao. Markovian scale prediction: A new era of visual autoregressive generation. *arXiv preprint arXiv:2511.23334*, 2025. 2

[68] Yinhan Zhang, Yue Ma, Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Magiccolor: Multi-instance sketch colorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15205–15217, 2025. 2

[69] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19513–19524, 2025. 2

[70] Yu Zhang, Jialei Zhou, Xinchen Li, Qi Zhang, Zhongwei Wan, Duoqian Miao, Changwei Wang, and Longbing Cao. Enhancing text-to-image diffusion transformer via split-text conditioning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[71] Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*, 2024. 2