GROUP DIFFUSION TRANSFORMERS ARE UNSUPERVISED MULTITASK LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

While large language models (LLMs) have revolutionized natural language processing with their task-agnostic capabilities, visual generation tasks such as image translation, style transfer, and character customization still rely heavily on supervised, task-specific datasets. In this work, we introduce **Group Diffusion Transformers (GDTs)**, a novel framework that unifies diverse visual generation tasks by redefining them as a **group generation** problem. In this approach, a set of related images is generated simultaneously, optionally conditioned on a subset of the group. GDTs build upon diffusion transformers with minimal architectural modifications by concatenating self-attention tokens across images. This allows the model to implicitly capture cross-image relationships (*e.g.*, identities, styles, layouts, surroundings, textures, and color schemes) through caption-based correlations. Our design enables scalable, unsupervised, and task-agnostic pretraining using extensive collections of image groups sourced from multimodal internet articles, image galleries, and video frames. We evaluate GDTs on a comprehensive benchmark featuring over 200 instructions across 30 distinct visual generation tasks, including picture book creation, font design, style transfer, sketching, colorization, drawing sequence generation, and character customization. Our models achieve competitive **zero-shot** performance without any additional fine-tuning or gradient updates. Furthermore, ablation studies confirm the effectiveness of key components such as data scaling, group size, and model design. These results demonstrate the potential of GDTs as scalable, general-purpose visual generation systems. We will release the code and models to support further research.

1 INTRODUCTION

The advent of large language models (LLMs) has brought a paradigm shift in natural language processing (NLP) Radford et al. (2019); Raffel et al. (2020); Brown (2020); Ouyang et al. (2022); Zhang et al. (2022); Touvron et al. (2023a;b); Dubey et al. (2024), enabling a wide range of tasks to be approached in a task-agnostic manner. These models, trained on vast corpora, can generate coherent and contextually relevant content across various domains without the need for task-specific fine-tuning, setting a new standard for what is achievable in NLP. However, this level of task generalization has yet to be fully realized in the field of visual generation. Unlike NLP, visual generation tasks – such as pose transfer Shen et al. (2023); Lu et al. (2024), image translation Ho et al. (2024); Rodatz et al. (2024), customization Jones et al. (2024); Wei et al. (2023), stylization Huang et al. (2024); Yang et al. (2023), and font creation Wang et al. (2023a); Yang et al. (2024) – remain largely siloed, relying heavily on supervised learning paradigms. These tasks often demand extensive task-specific datasets and additional modules, such as LoRAs Jones et al. (2024); Smith et al. (2023); Luo et al. (2023), adapters Ye et al. (2023a); Mou et al. (2024), visual encoders Giannone et al. (2022); Kumar et al. (2024); Xu et al. (2024), and ControlNets Zhang et al. (2023); Zhao et al. (2024), to achieve satisfactory performance.

This reliance on specialized data and architectures presents significant challenges for scalability and generalization. First, it limits scalability by failing to leverage the vast amount of weakly supervised data available on the Internet; creating and curating task-specific datasets is human-laboring. Second, it restricts models' adaptability to unseen tasks. Third, cross-task adaptation is lacking, particularly in compositional control, where multiple tasks are implicitly managed. For example, consider creating a picture book Jin & Song (2023); Wang et al. (2023b), characters, environments,

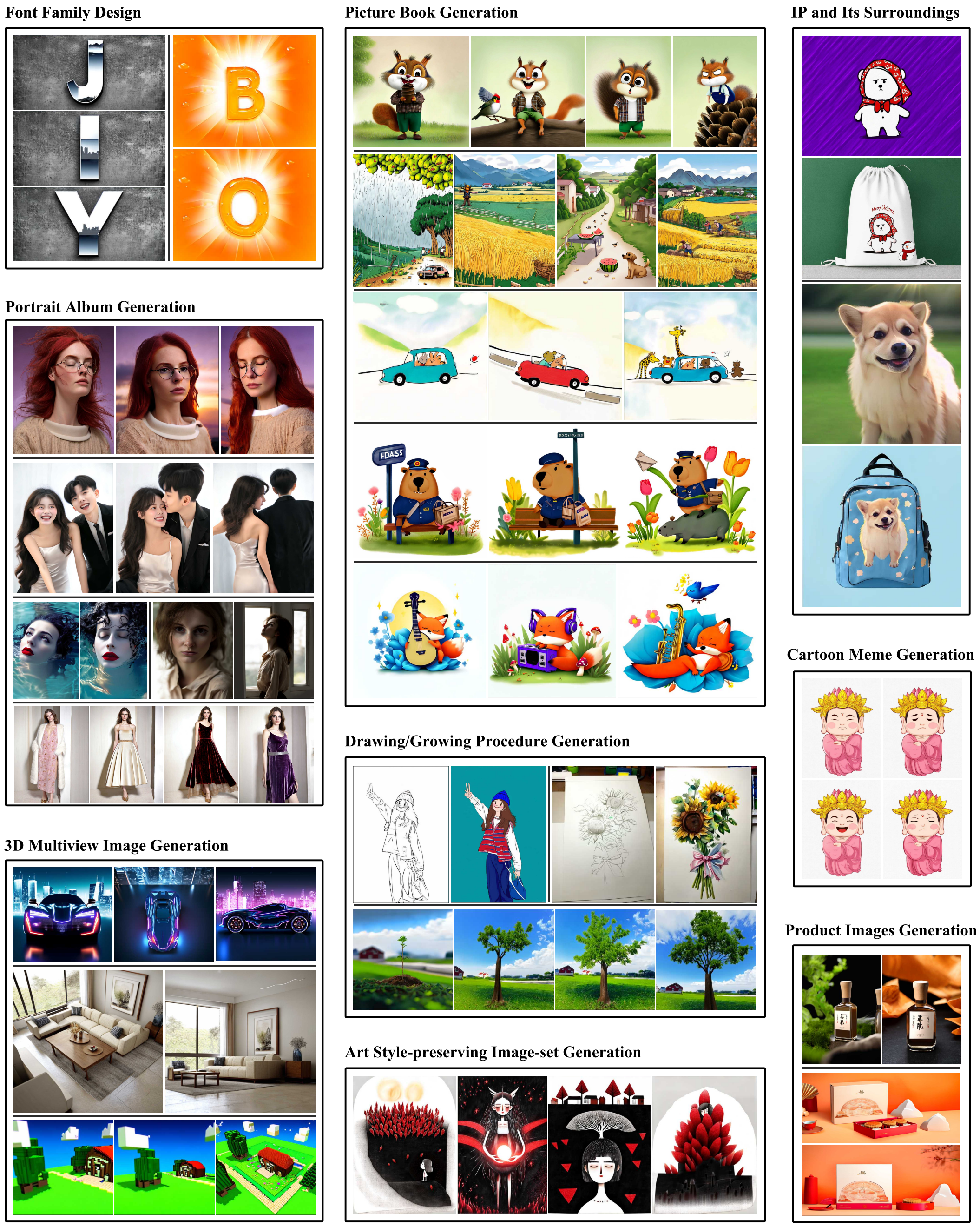


Figure 1: **Group Diffusion Transformers perform a vast array of visual generation tasks in a unified framework termed group generation.** Note that **NO** task-specific dataset and **NO** additional gradient update is applied. The model is automatically generalized to these tasks after unsupervised training on image groups. For simplicity, textual descriptions of images are omitted here, which can be found in Appendix.

and attire must be dynamically adjusted, requiring decisions on which elements to keep consistent and which to vary. Finally, we hypothesize that training on single-task, shallow-domain datasets leads to the lack of generalization in real-world applications. To truly unlock the potential of visual generation, it is crucial to develop models capable of performing a wide range of tasks in a task-agnostic manner. This demands a shift in how we conceptualize and approach these tasks.

Our key insight is that most, *if not all*, visual generation tasks can be reformulated within a unified framework that we term the **group generation** problem. In this framework, the objective is

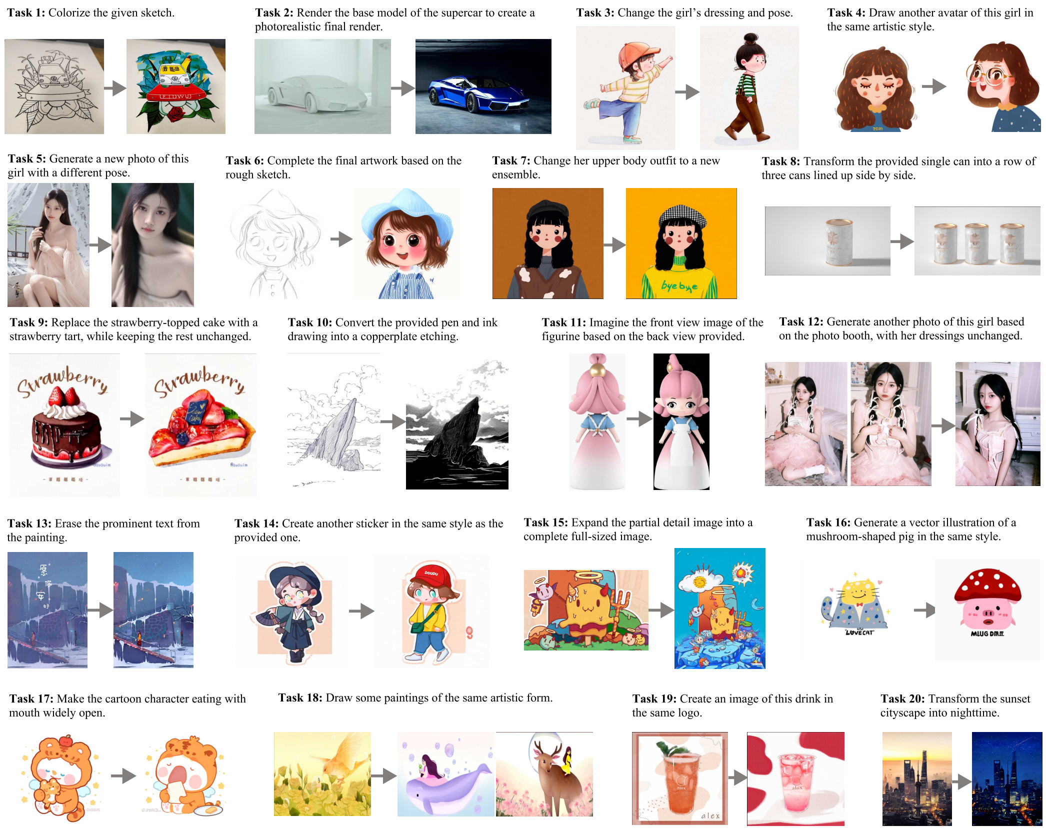


Figure 2: **When conditioned on a subset of the group data, Group Diffusion Transformers could perform conditional group generation in the inpainting fashion.** Note that the model is automatically generalized to these tasks after unsupervised training on image groups. Textual descriptions of images are omitted here (can be found in Appendix), and we summarize them into brief task descriptions.

to generate a set of correlated data, or a *group*, optionally conditioned on a subset of this group. For instance, tasks such as generating picture books Jin & Song (2023); Wang et al. (2023b), font images Wang et al. (2023a); Yang et al. (2024), or emoticons Mittal et al. (2020) involve producing multiple images with distinct yet related descriptions simultaneously. The inherent correlations are implicitly captured through the relationships among these descriptions. Conversely, tasks like sketching Voynov et al. (2023); Wang et al. (2023c), colorization Zabari et al. (2023); Carrillo et al. (2023); Liang et al. (2024), character-specific image generation Zdenek & Nakayama (2023); Kou et al. (2023), and multiview image generation from a single image Liu et al. (2023b); Shi et al. (2023) can be framed as conditional group generation problems, where a subset of the group data is provided as a reference. Figure 1 and 2 provide examples of group generation and conditional group generation. By reframing these tasks as group generation problems, we leverage the power of unsupervised learning to address a broad spectrum of tasks without the need for task-specific supervision, simplifying the learning process and broadening applicability.

One of the most compelling advantages of the **group generation** framework is its natural alignment with the vast amount of data available on the Internet. Multimodal articles, image galleries, and multi-shot videos are just a few examples of readily accessible sources of group data. Each of these sources inherently captures the relationships between different data elements, offering a form of free supervision that is both scalable and diverse. The availability of such abundant group data not only reduces the need for labor-intensive data annotation but also enables the training of models on a wide array of tasks simultaneously, further enhancing generalizability.

To address the group generation problem, we introduce a minimalistic modification to diffusion transformers Peebles & Xie (2023); Esser et al. (2024a); Chen et al. (2023a), termed **Group Diffusion Transformers (GDTs)**. The core idea is to concatenate self-attention tokens across a group of inputs, allowing the model to learn the correlations and variations within the group. This modification is straightforward, requiring minimal changes to the underlying architecture of diffusion transformers (DiTs), yet it significantly enhances the model's ability to capture relationships among multiple generated data. To address reference-based generation problems, such as style transfer Huang et al. (2024); Yang et al. (2023) and image translation Ho et al. (2024); Rodatz et al. (2024), we incorporate techniques like SDEdit Meng et al. (2021) and inpainting Xie et al. (2023); Xu et al. (2024). These methods enable the model to generate the remaining elements of a group when conditioned on a subset of inputs. Figure 3 provides a detailed architectural overview of GDTs. The straightforward design of GDTs makes it easy to implement and shows promise for efficient scaling.

To evaluate the capabilities of our model, we first introduce a user interface that can automatically convert user instructions into textual descriptions of the target image group to support group generation. Then, we construct a comprehensive benchmark that covers a wide range of visual generation tasks, both with and without reference images. All tasks are performed in a zero-shot setting, without any parameter or architectural modifications. Despite the absence of task-specific supervision during training, our model demonstrates promising performance across most tasks. Finally, we conduct ablation studies to examine the impact of key components in our framework, such as data scale, group size, model design and quality tuning, on overall performance.

2 Approach

The core of our approach is to reformulate visual generation tasks into a *group generation* problem and solve it using minimally modified diffusion transformers. We begin by detailing how these tasks are reformulated, followed by a comprehensive introduction to our model, its architecture, the data employed, the training procedure, and the inference stage.

2.1 Problem Formulation

We propose that a vast array of visual generation tasks can be unified under a single framework we term the **group generation** problem. In this framework, the objective is to generate a group of n elements $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, where each element is conditioned on its respective context (*e.g., image descriptions*) $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n\}$. The relationships among these elements are implicitly defined by the interdependencies within their contextual conditions. Optionally, a subset of $0 \leq m < n$ elements of $\mathbf{x}$ can be provided as reference data, with the task being to generate the remaining $(n-m)$ elements. This formulation naturally encapsulates a variety of tasks:

- **Text-to-Image:** A special case where the group size $n = 1$ and the reference subset size $m = 0$. The task is to generate a single image from a textual description.
- **Font Generation:** Here, the group size $n > 1$ corresponds to the number of characters to generate, with $m = 0$.
- **Picture Book Generation:** Similar to font generation, the group size $n > 1$ corresponds to the number of picture book pages, with $m = 0$. The descriptions capture the connections and variations across the pages.
- **Identity Preservation:** Here, the group size $n > 1$ corresponds to the number of photos with the same identities to generate, with $m = 0$. Identity-specific information is reflected in the descriptions, such as names or other identifiers.
- **Local Editing:** In this task, the group size is $n = 2$ with a reference subset size $m = 1$. One reference image is provided, and the model generates the edited image based on the differences captured in their descriptions.
- **Image Translation:** Similarly, the group size is $n = 2$ with a reference subset size $m = 1$. A reference image from one domain is converted to another domain according to their descriptions.
- **Subject Customization:** The task involves generating $(n-m) \geq 1$ images, where $1 \leq m < n$ character images are used as references.
- **Style Adaptation:** In this task, $(n-m) \geq 1$ corresponds to the number of stylized images to be generated, with $m = 1$ being the reference image guiding the target style.

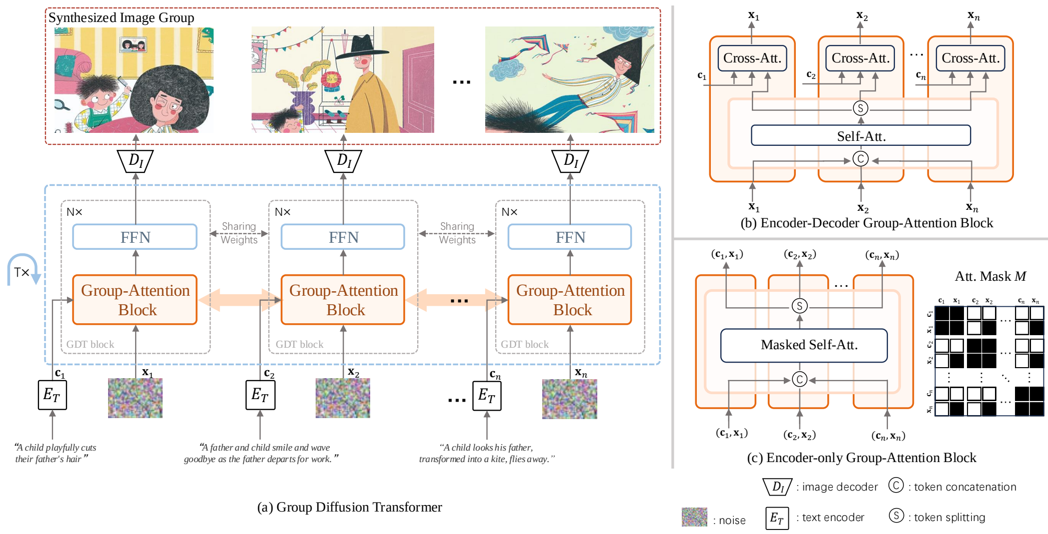


Figure 3: **The overview of Group Diffusion Transformer, which takes minimal adaptations for the encoder-decoder and encoder-only visual generation architectures.** We make a straightforward modification on self-attention blocks by concatenating image tokens across group inputs, allowing to learn inter-image correlations.

These examples illustrate just a few of the many tasks that can be naturally expressed within the *group generation* framework. Across these tasks, the task hints are naturally embedded within the group element descriptions, much like how a human might communicate with a designer. This unified framework simplifies the approach to diverse visual generation tasks and paves the way for scalable, generalized solutions.

2.2 MODEL AND ARCHITECTURES

To tackle the group generation problem, it is crucial to establish connections between multiple group elements during the generation process, allowing the model to perceive and utilize the correlations among these elements. Our approach involves a straightforward modification: concatenating tokens across group inputs within the self-attention blocks of diffusion transformers. This enables tokens from different data elements to interact with one another throughout the model's layers.

For different text-conditioned visual generation architectures, we make minimal adaptations to accommodate our approach:

- **Encoder-Decoder:** In architectures like PixArt Chen et al. (2023a), each transformer block includes a self-attention operation for the image, cross-attention for interaction between image and text, and a feed-forward network. We choose to concatenate all the image tokens in self-attention blocks, which allows every token attends to all the image tokens within the group. After self-attention operation, concatenated image tokens are split back correspondingly. Then, in cross-attention blocks, each image token attends only to the text embeddings associated with its respective description. This setup is illustrated in Figure 3 (b).
- **Encoder-Only:** Examples like Stable Diffusion 3 Esser et al. (2024a) and FLUX Labs (2024) feature transformer blocks with self-attention blocks and feed-forward networks. We modify the self-attention operation into a masked version, which is depicted in Figure 3 (c). Specifically, image tokens $\mathbf{x}_i$ as well as text tokens $\mathbf{c}_i$ are first concatenated with each other all over the group. Then, we calculate the masked self-attention, where the mask is designed for allowing every image token attends to all tokens across the group while allowing context tokens only attend to image tokens as well as themselves. Concretely, let $M(\mathbf{a}_j, \mathbf{b}_k)$ indicate the attention mask for tokens in $\mathbf{a}_j$ and $\mathbf{b}_k$, where $\mathbf{a}, \mathbf{b} \in \{\mathbf{c}, \mathbf{x}\}, 0 \leq j, k \leq n$. Then, $M(\mathbf{a}_j, \mathbf{b}_k)$ is decided by

$$M(\mathbf{a}_j, \mathbf{b}_k) = \begin{cases} 1 & \text{if } (j = k) \text{ or } (\mathbf{a} \in \mathbf{x} \text{ and } \mathbf{b} \in \mathbf{x}) \\ 0 & \text{else} \end{cases}. \quad (1)$$

2.3 Training Dataset

We focus on image-related tasks in this work, which requires a high-quality, large-scale, and diverse image group dataset. While existing multimodal datasets like MINT-1T Awadalla et al. (2024) are large, they fall short of our pretraining needs due to low image quality and biased group type distribution relative to real-world visual generation applications. Thus, we construct our own dataset by sourcing image groups from multimodal Internet articles.

Our dataset creation process involve several key steps: (1) We collect a substantial amount of multimodal data, extracting images while preserving their original order to maintain group integrity. (2) A small subset of these image groups is manually annotated as either positive (suitable for retention) or negative (to be discarded). (3) Using these annotations, we train a binary classifier to score and filter the collected image groups. (4) We perform deduplication across and within groups to eliminate redundant groups and images. After preprocessing, we compile a dataset of approximately 500,000 image groups, with the distribution of group size illustrated in Figure 4.

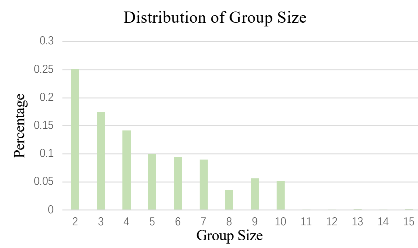


Figure 4: **Distribution of group size in our training dataset.**

The next crucial step is to generate descriptions that accurately capture the correlations among the images within each group. To achieve this, we utilize our internal multimodal large language models (MLLMs), iteratively testing and refining prompts to ensure the generated descriptions are stable and applicable across different group types. In Figure 5, we show the prompt we used, as well as the resulting group image descriptions.

While pretraining on our large-scale dataset provides a solid foundation for learning correlations with Group Diffusion Transformers (GDTs), it is common practice in visual generation tasks to conduct a supervised fine-tuning stage to enhance generation details and aesthetics. To this end, we curate a smaller, high-quality subset of approximately 10,000 image groups. These groups were selected for their strong correlations, high image quality, aesthetic appeal, and diversity. Fine-tuning our pretrained models on this curated dataset significantly improves both the image quality and content consistency in group generation, where the comparison can be found in Section 4.2.5.

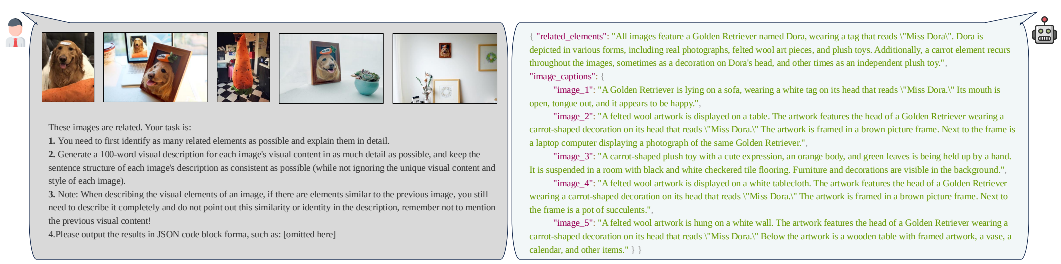


Figure 5: **Example of our training dataset, where the group images are captioned through prompting our internal MLLMs.**

2.4 Training Process

We initialize the Group Diffusion Transformers (GDTs) with weights from pre-trained text-to-image models, such as PixArt-α Chen et al. (2023a) and Stable Diffusion 3 Peebles & Xie (2023). Since GDTs introduce no additional parameter to the existing diffusion transformers, the pretrained weights are fully compatible. During both pretraining and supervised fine-tuning, we uniformly sample group sizes ranging from 1 to 4, dynamically adjusting the batch size to maintain consistent GPU memory usage. This approach ensures balanced performance across different group sizes. The model undergoes pretraining for approximately 100,000 steps, followed by fine-tuning on a curated dataset for around 5,000 steps. All training is conducted on A100 GPUs. We adopt the same hyperparameter settings as the official models in PixArt-α and Stable Diffusion 3.

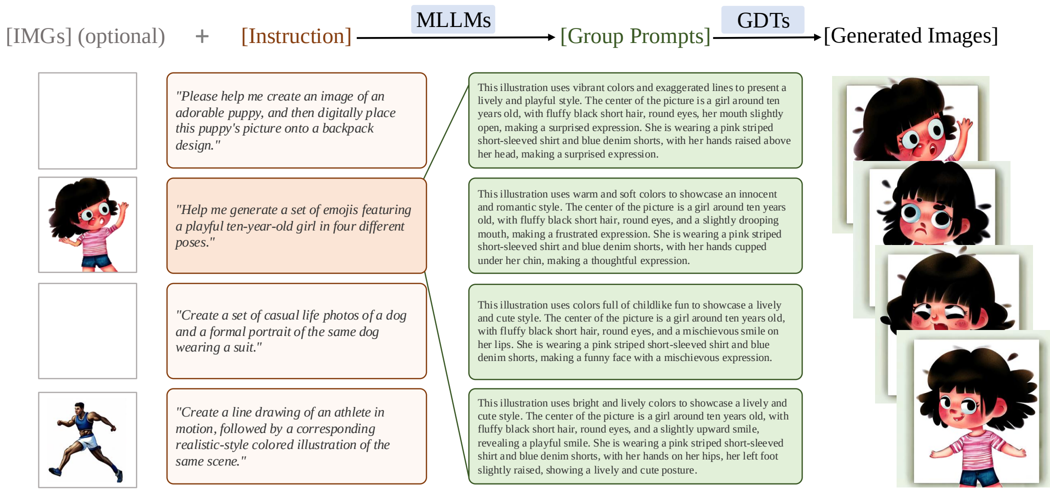


Figure 7: **We build a user interface that automatically converts the user instruction into group prompts using MLLMs, which is useful in the inference stage of GDTs.**

2.5 User Interface

Considering it is tedious to write a group of prompts in the inference stage, we build a **user interface** to provide a convenient interaction with the GDTs. As illustrated in Figure 7, we follow the pipeline of [`Instruction`] → [`Group Prompts`] → [`Generated Images`] for group generation, and [`IMGs`] + [`Instruction`] → [`Group Prompts`] → [`Generated Images`] for conditional group generation. Specifically, we leverage MLLMs to convert the user instruction into group prompts, where the MLLM could analyze the number of group prompts and the corresponding tasks. For example, if the instruction is "Draw a line sketch of a female character and the corresponding colored photo", the MLLM can deduce that this instruction should be transformed into two prompts, categorizing the task as sketch coloring.

3 Benchmark

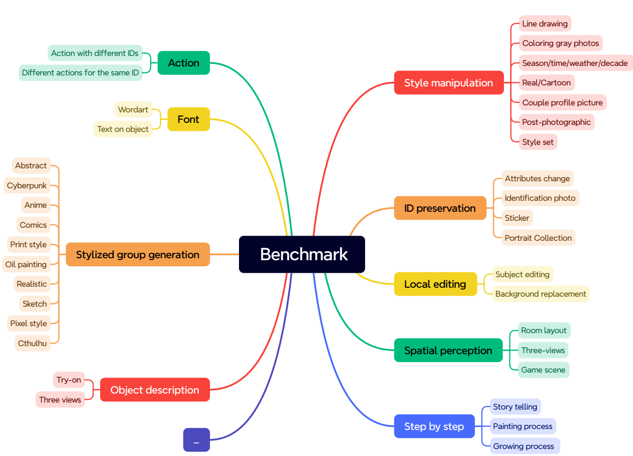
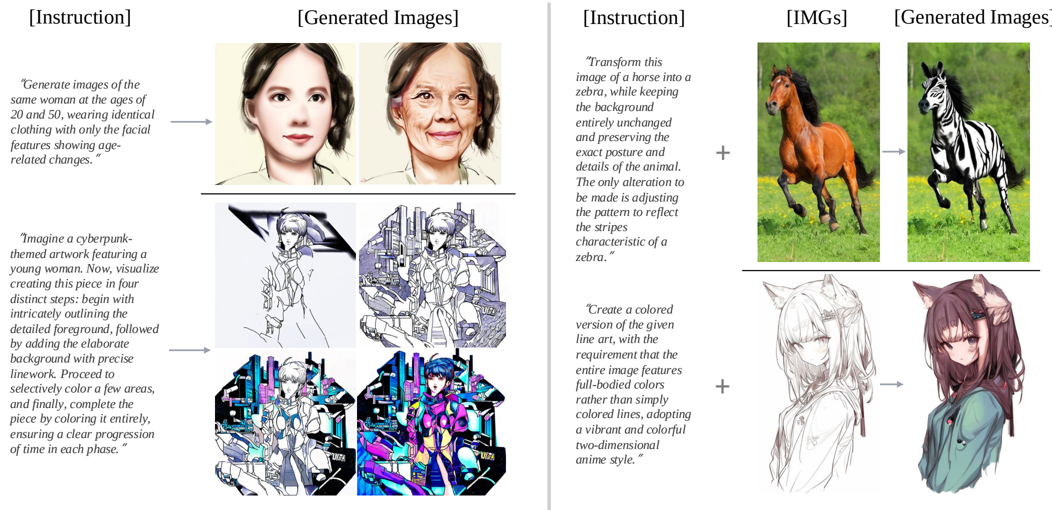


Figure 6: **Overview of our benchmark, covering about 30 distinct types of generation tasks.**

Given the diverse nature of visual generation tasks, evaluating the performance of our Group Diffusion Transformers (GDTs) presents unique challenges. Therefore, we design a benchmark that spans a wide array of tasks as shown in Figure 6. Specifically, our benchmark consists of over 200 instructions, each corresponding to one of 30 distinct types of visual generation tasks. This diversity enables a thorough assessment of the generalization capabilities of GDTs across various scenarios.

This evaluation suit encompasses tasks such as identity preservation, local editing, subject customization, font generation, and stylized group generation. Among these coarse-grained categories, further fine-grained tasks are expanded. For example, step-by-step generation contains subtasks like story telling Zhou et al. (2024), painting process Song et al. (2024), and growth process. Besides, all the textual descriptions in this benchmark are created through our user interface.



(a) Group generation (b) Conditional group generation

Figure 8: **Generated results of GDTs on our benchmark, including group generation and conditional group generation.**

4 RESULTS

4.1 USER STUDY

We first qualitatively evaluate the generated results of GDTs on our proposed benchmark as shown in Figure 8. GDTs could perform both group generation and conditional group generation according to the user instructions. Note that the task scope of this benchmark is effectively limited by our imagination, but thanks to our unsupervised and task-agnostic pretraining, GDTs can theoretically be generalized to *arbitrary* visual generation tasks.

Table 1: **User study on our benchmark.** Human evaluation on three questions in a five-point scale.

Models	Q1	Q2	Q3
group generation			
PixArt-α	3.44	3.89	3.78
Stable Diffusion 3	3.20	3.35	3.29
conditional group generation			
PixArt-α	3.15	3.56	3.68
Stable Diffusion 3	3.02	3.27	3.34

In our user study, we mainly adopt human ratings to assess the performance of GDTs on the benchmark. Three questions are included to measure the prompt following ability, content consistency within the image group, and the overall instruction following ability, namely: **Q1: Prompt following on each image within the group**: **Q2: Content consistency among generated group images, regardless of prompts**, **Q3: Instruction following on the generated group images.** Evaluators are asked to rate on three questions in the scale from 1 to 5, where 5 signifies perfection and 1 denotes the lowest quality. The final evaluation score is derived from the average ratings across all tasks, which serves as a robust indicator of the overall performance and its potential for real-world applications. The human-rated results are illustrated in Table 1, where GDTs achieve overall satisfaction (higher than 3) on all of the three questions.

4.2 ABLATION ANALYSIS

4.2.1 METRICS

While our benchmark with over 200 instructions could well evaluate model's capabilities on a five-point scale, we would like to compare these ablated models in a more nuanced and quantitative manner in our ablation experiments. Therefore, we mainly present the objective metrics like FID and CLIP score. To be specific, we measure image fidelity by calculating FID on the validation set using 50k images. We assess content consistency and prompt adherence within each group by averaging CLIP similarities across every image-image and image-text pairs, respectively. In terms of reference-based generation, we adopt the same metrics but exclude pairs that involve the reference images themselves, as well as pairs between reference images and their corresponding texts.

Table 2: **Performance evaluation on key components of GDTs.** We investigate the impacts of data scale, group size, model design, and quality tuning on encoder-decoder and encoder-only models.

Settings	PixArt-α (Encoder-Decoder)			Stable Diffusion 3 (Encoder-Only)		
	FID-50k	Content Consistency	Prompt Adherence	FID-50k	Content Consistency	Prompt Adherence
Data Scaling						
5k groups	8.40	0.747	0.291	8.95	0.740	0.298
50k groups	12.06	0.767	0.293	10.92	0.760	0.302
500k groups	15.91	0.778	0.300	11.30	0.761	0.305
Group Size						
groupsize = 2	15.69	0.784	0.299	12.37	0.763	0.301
groupsize = 4	18.19	0.761	0.291	13.85	0.739	0.298
groupsize = 8	48.26	0.701	0.252	18.28	0.701	0.290
Inpainting						
SDEdit	15.71	0.702	0.299	12.15	0.751	0.303
trainable	10.91	0.725	0.287	10.94	0.755	0.298
Quality Tuning						
before	15.91	0.778	0.300	11.30	0.761	0.305
after	12.53	0.792	0.298	10.03	0.781	0.303

4.2.2 DATA SCALING

Without the demand of task-specific supervision, it is quite easy to acquire a large abundance of group data from the Internet. We scale the training data to 5k, 50k, and 500k groups, to explore the impact of data scale in GDTs. As illustrated in Table 2, with the increase of the amount of training data, GDTs behave increasingly better in content consistency and prompt adherence. Interestingly, we find that FID would become lower when training on less data, which may be that it is easier to overfit to small datasets. We plan to further scale up our data to the level of hundreds of millions of groups in the future, in order to fully leverage the potential of GDTs.

4.2.3 GROUP SIZE

We gradually increase the upper limit of group size to 2, 4, and 8, and perform inference based on that limit. Note that doubling the group size will, in turn, double the sequence length in self-attention, leading to a corresponding increase in computational complexity, so we cap the maximum group size at 8 in our ablation. From the ablated results in Table 2, we find that larger group sizes lead to a more pronounced performance decline in image quality, content consistency, and prompt adherence. The reason may be that it is more difficult to learn the complex relationships across a large group of images. Besides, the scarcity of data of large group sizes prevents the model from being adequately trained. In the future, we would greatly scale our training data.

4.2.4 SDEDIT OR INPAINTING

When conditioned on a subset of the group data, using methods like SDEdit Meng et al. (2021) or trainable inpainting Xie et al. (2023); Xu et al. (2024), GDTs can be instructed to generate the remaining data of the group. Specifically, SDEdit is a training-free technique which provides the reference images that are added with the same noise step as the generated images during the denoising stage. In contrast, trainable inpainting concatenates the reference image to the noised one in channel dimension, allowing the model to "copy" the reference images and generate the remaining ones. In our ablation study, as illustrated in Table 2, it is observed that trainable inpainting performs better in image quality and content consistency, while the training-free SDEdit is good at prompt adherence. We adopt the model design of trainable inpainting in our GDTs.

4.2.5 QUALITY TUNING

While quality tuning is a common practice in visual generation models to enhance aesthetic appeal, we investigate its impact under the paradigm of group generation. As illustrated in Table 2, after the supervised fine-tuning on a small subset of high-quality image groups, GDTs exhibit significantly better image quality. We also find that quality tuning helps generating image groups with higher content consistency, while barely compromising the adherence to textual descriptions.

5 Related Work

5.1 Text-to-Image Generation

The emergence of DDPM Ho et al. (2020) has catalyzed rapid advancements in text-to-image (T2I) generation. Earlier frameworks focused on T2I generation in pixel space, exemplified by GLIDE Nichol et al. (2022) and Imagen Saharia et al. (2022). In contrast, Stable Diffusion Rombach et al. (2022) introduced latent space for T2I generation, while DALLE-2 (unCLIP)Ramesh et al. (2022a) expanded this to a multimodal latent space. EMUDai et al. (2023) demonstrated that supervised fine-tuning on a small set of appealing images can significantly enhance generation quality. Unlike U-Net architectures, several approaches, including DiT Peebles & Xie (2023), Pixart Chen et al. (2023a), HunyuanDiT Li et al. (2024b), and SD3 Esser et al. (2024b), adopt transformers as their backbone.

5.2 Controllable Text-to-Image Generation

Personalization. Personalization in T2I generation Cui et al. (2024); Salehi et al. (2024); Ham et al. (2024); Wang et al. (2024) aims to capture concepts like subject Li et al. (2023a); Kumari et al. (2023), person Xiao et al. (2023); Li et al. (2024a); Chen et al. (2024b; 2023b), style Liu et al. (2023a); Sohn et al. (2023), and image Ye et al. (2023b); Xu et al. (2023); Ramesh et al. (2022b). Techniques like Textual Inversion Gal et al. (2022) and DreamBooth Ruiz et al. (2022) facilitate concept embedding. Subject-driven methods Valevski et al. (2023); Chen et al. (2024b) use face recognition models for personalization.

Spatial Control. Spatial control in T2I generation Li et al. (2023b) is crucial for representing image structure. ControlNet Zhang et al. (2023) and UniControl Qin et al. (2023) are examples of models that incorporate positional signals for spatial control.

Advanced Controllable Text-to-Image Generation. New directions in controllable T2I generation include Attend-and-Excite Chefer et al. (2023), Composer Huang et al. (2023), Cocktail Hu et al. (2023), Cones Liu et al. (2023c), Universal Guidance Bansal et al. (2023), EMU2 Sun et al. (2024), and FreeDom Yu et al. (2023), which aim to enhance text alignment and achieve universal control.

5.3 Generalization Ability of Generative Models

Beyond fundamental generative capabilities, recent approaches are investigating the generalization and versatility of models. ControlNeXt Peng et al. (2024) is designed to support both images and videos while incorporating diverse forms of control information. EMU2 Sun et al. (2024) demonstrates task-agnostic in-context learning capabilities. MT-Diffusion Chen et al. (2024a) achieves multi-modality diffusion through multi-task learning.

In contrast to the aforementioned methods, Group Diffusion Transformers aim to provide a general-purpose visual generation framework with the following capabilities: 1) no need for task-specific pretraining or finetuning; 2) generating multiple images in parallel; 3) conditioning on text or images; and 4) enabling zero-shot task generalization.

6 Conclusion and Limitations

We reformulate most visual generation tasks into a **group generation** problem, thereby introducing a unified framework named **Group Diffusion Transformers** (GDTs). We present that with scalable, unsupervised, and task-agnostic pretraining on group data, GDTs could achieve competitive zero-shot performance on a vast array of visual generation tasks. Our results demonstrate the potential of GDTs as scalable, general-purpose visual generation systems.

Moreoever, we point out that there is still a discrepancy in image quality between GDTs and the state-of-the-art text-to-image models. The amount of group data for pretraining is also not sufficient yet, which has not fully unleashed the model's capabilities. We are optimistic that with an enlarged group dataset, we can further optimize the model's performance and reduce the discrepancy. In the future, we also plan to extend the time dimension of GDTs to enable multi-shot video generation, which can be naturally expressed under our group generation framework.

REFERENCES

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *arXiv preprint arXiv:2406.11271*, 2024.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023. URL https://arxiv.org/abs/2302.07121.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Hernan Carrillo, Michaël Clément, Aurélie Bugeau, and Edgar Simo-Serra. Diffusart: Enhancing line art colorization with conditional diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3486–3490, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42:1–10, 07 2023.

Changyou Chen, Han Ding, Bunyamin Sisman, Yi Xu, Ouye Xie, Benjamin Z. Yao, Son Dinh Tran, and Belinda Zeng. Diffusion models for multi-task generative modeling. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=cbv0sBIZh9.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ArXiv*, abs/2310.00426, 2023a. URL https://api.semanticscholar.org/CorpusID:263334265.

Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models, 2023b.

Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, and Zhendong Mao. Dreamidentity: Enhanced editability for efficient face-identity preserved image generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2):1281–1289, Mar. 2024b.

Siying Cui, Jia Guo, Xiang An, Jiankang Deng, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models, 2024. URL https://arxiv.org/abs/2403.13535.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack, 2023. URL https://arxiv.org/abs/2309.15807.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206, 2024a. URL https://api.semanticscholar.org/CorpusID:268247980.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024b. URL https://arxiv.org/abs/2403.03206.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2022.

Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhang, and Tobias Hinz. Personalized residuals for concept-driven text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8186–8195, June 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Ming-Yang Ho, Che-Ming Wu, Min-Sheng Wu, and Yufeng Jane Tseng. Every pixel has its moments: Ultra-high-resolution unpaired image-to-image translation via dense normalization. *arXiv preprint arXiv:2407.04245*, 2024.

Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. Cocktail: Mixing multi-modality controls for text-conditional image generation. *arXiv*, 2023.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Ze Jin and Zorina Song. Generating coherent comic with rich story using chatgpt and stable diffusion. *arXiv preprint arXiv:2305.11067*, 2023.

Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. *arXiv preprint arXiv:2405.01536*, 2024.

Ziyi Kou, Shichao Pei, Yijun Tian, and Xiangliang Zhang. Character as pixels: A controllable prompt adversarial attacking framework for black-box text guided image generation models. In *IJCAI*, pp. 983–990, 2023.

Manoj Kumar, Neil Houlsby, and Emiel Hoogeboom. Semantica: An adaptable image-conditioned diffusion model. *arXiv preprint arXiv:2405.14857*, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023.

Black Forest Labs. Flux.1, 2024. URL https://github.com/black-forest-labs/flux.

Dongxu Li, Junnan Li, and Steven Hoi. BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=g6We1SwaY9.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521, June 2023b.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024b. URL https://arxiv.org/abs/2405.08748.

Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color: Multimodal diffusion-based interactive image colorization. *arXiv preprint arXiv:2402.10855*, 2024.

Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023a.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023b.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *International Conference on Machine Learning*, 2023c. URL https://api.semanticscholar.org/CorpusID:257427549.

Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6420–6429, 2024.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://api.semanticscholar.org/CorpusID:245704504.

Paritosh Mittal, Kunal Aggarwal, Pragya Paramita Sahu, Vishal Vatsalya, Soumyajit Mitra, Vikrant Singh, Viswanath Veera, and Shankar M Venkatesan. Photo-realistic emoticon generation using multi-modal input. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 254–258, 2020.

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*, volume 38, pp. 4296–4304, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022a. URL `https://api.semanticscholar.org/CorpusID:248097655`.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022b. URL `https://arxiv.org/abs/2204.06125`.

Benjamin Rodatz, Ian Fan, Tuomas Laakkonen, Neil John Ortega, Thomas Hoffman, and Vincent Wang-Mascianica. A pattern language for machine learning tasks. *arXiv preprint arXiv:2407.02424*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. ViPer: Visual personalization of generative models via individual preference learning. *ECCV*, 2024.

Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*, 2023.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

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, 2024. URL `https://arxiv.org/abs/2406.06062`.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, 2023.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

Chi Wang, Min Zhou, Tiezheng Ge, Yuning Jiang, Hujun Bao, and Weiwei Xu. Cf-font: Content fusion for few-shot font generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1858–1867, 2023a.

Dejiang Wang, Zhuoran Zhai, Ngai Cheong, and Li Peng. Script-generated picture book technology based on large language models and aigc. In *Proceedings of the 7th International Conference on Digital Technology in Education*, pp. 104–108, 2023b.

Qiang Wang, Di Kong, Fengyin Lin, and Yonggang Qi. Diffsketching: Sketch control image synthesis with diffusion models. *arXiv preprint arXiv:2305.18812*, 2023c.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv*, 2023.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22428–22437, 2023.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7754–7765, October 2023.

Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8682–8692, 2024.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023.

Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2024.

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*, 2023a.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023b.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Jan Zdenek and Hideki Nakayama. Handwritten text generation with character-specific encoding for style imitation. In *International Conference on Document Analysis and Recognition*, pp. 313–329. Springer, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. 2024.

A Appendix

Figure 1: **Detailed results of Group Diffusion Transformers.**

Figure 2: **Detailed results of Group Diffusion Transformers.**

Figure 3: **Detailed results of Group Diffusion Transformers.**

Figure 4: **Detailed results of Group Diffusion Transformers.**

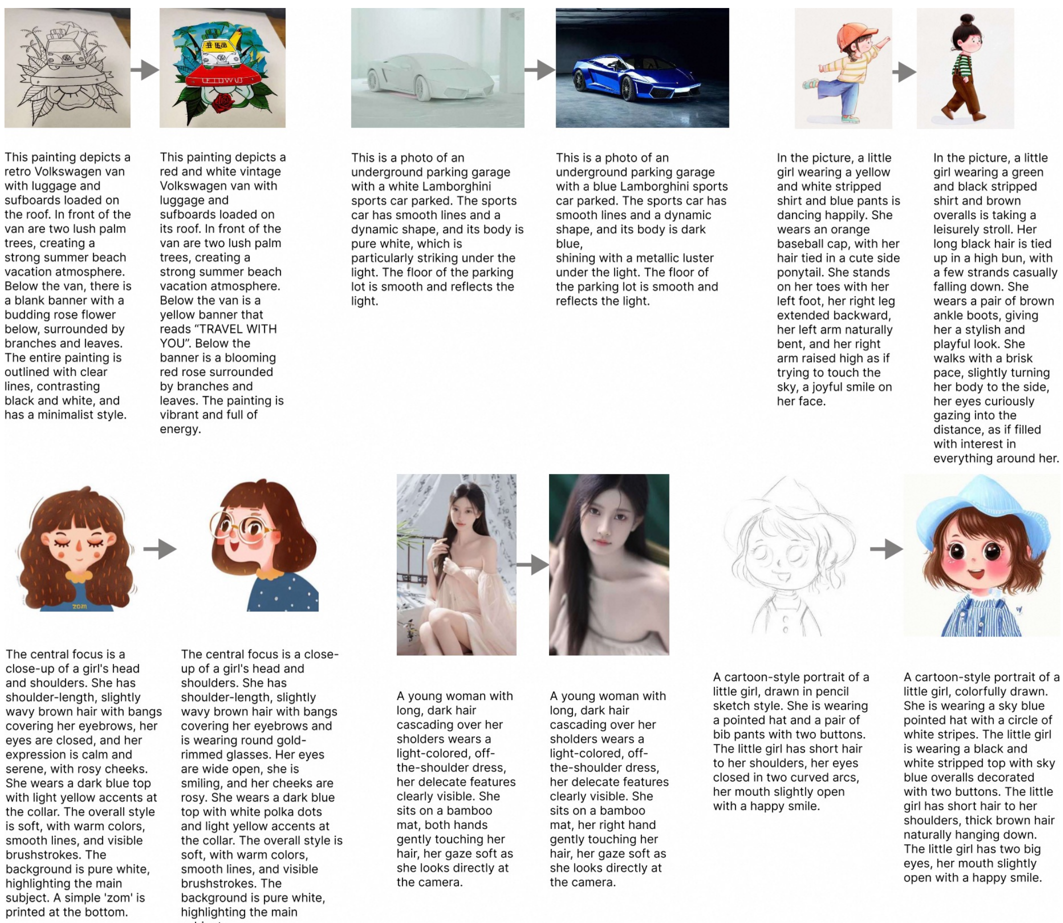


Figure 5: **Detailed results of Group Diffusion Transformers.**

Figure 6: **Detailed results of Group Diffusion Transformers.**

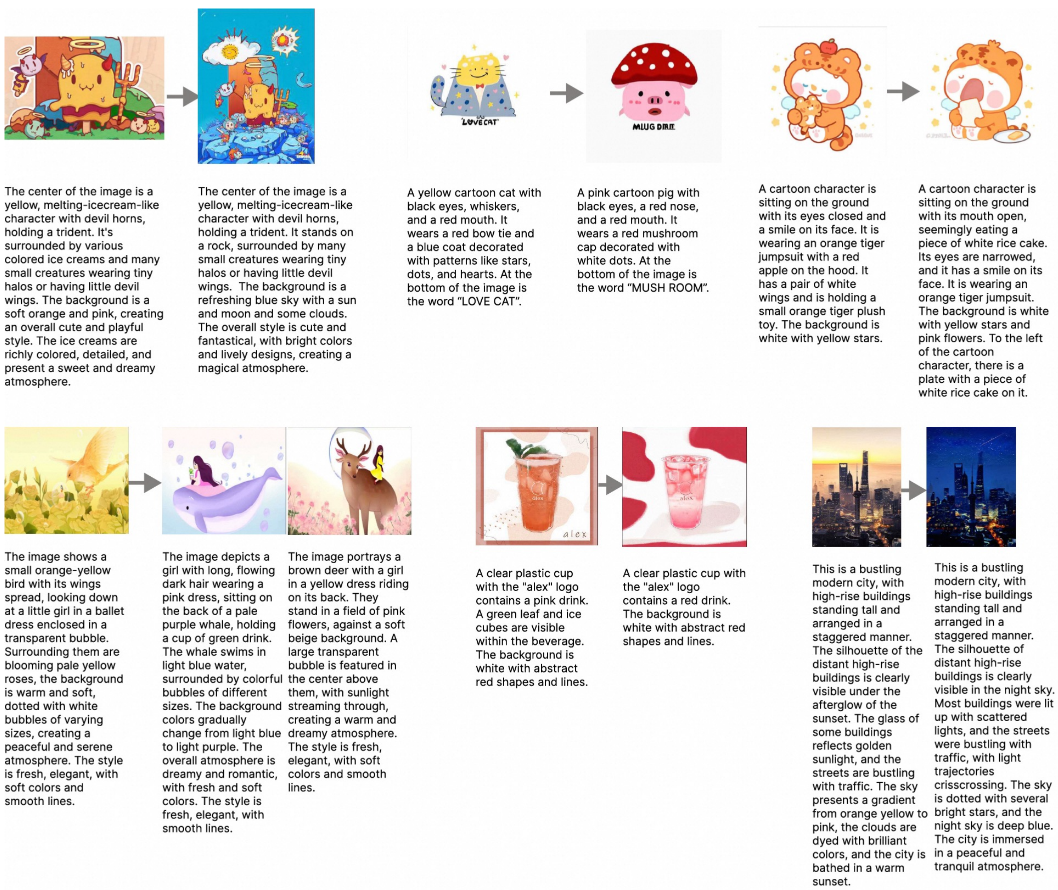


Figure 7: **Detailed results of Group Diffusion Transformers.**