

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,

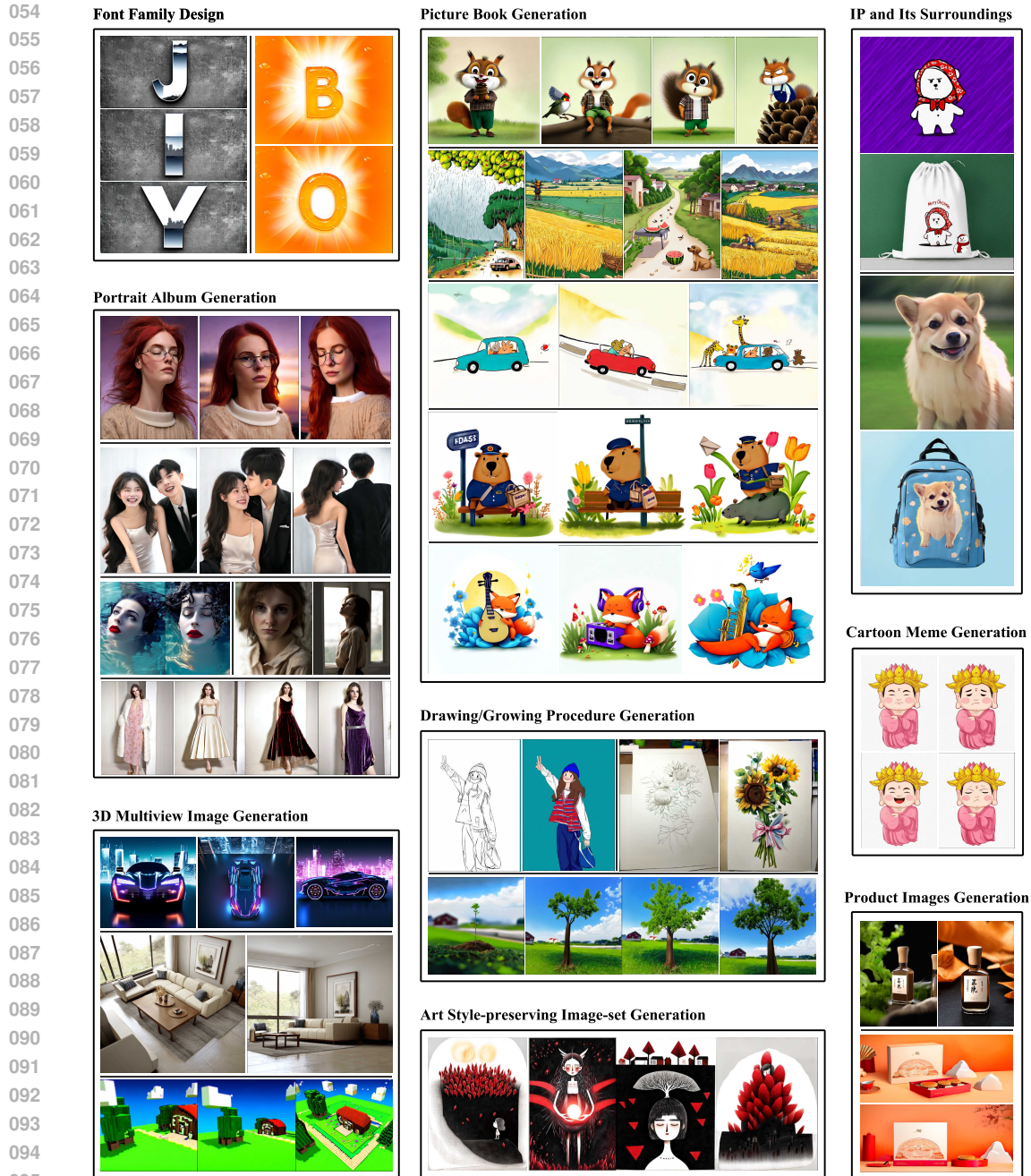


Figure 1: **Group Diffusion Transformers perform a vast array of visual generation tasks in a unified framework termed group generation.** Note that **NO** task-specific dataset and **NO** additional gradient update is applied. The model is automatically generalized to these tasks after unsupervised 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

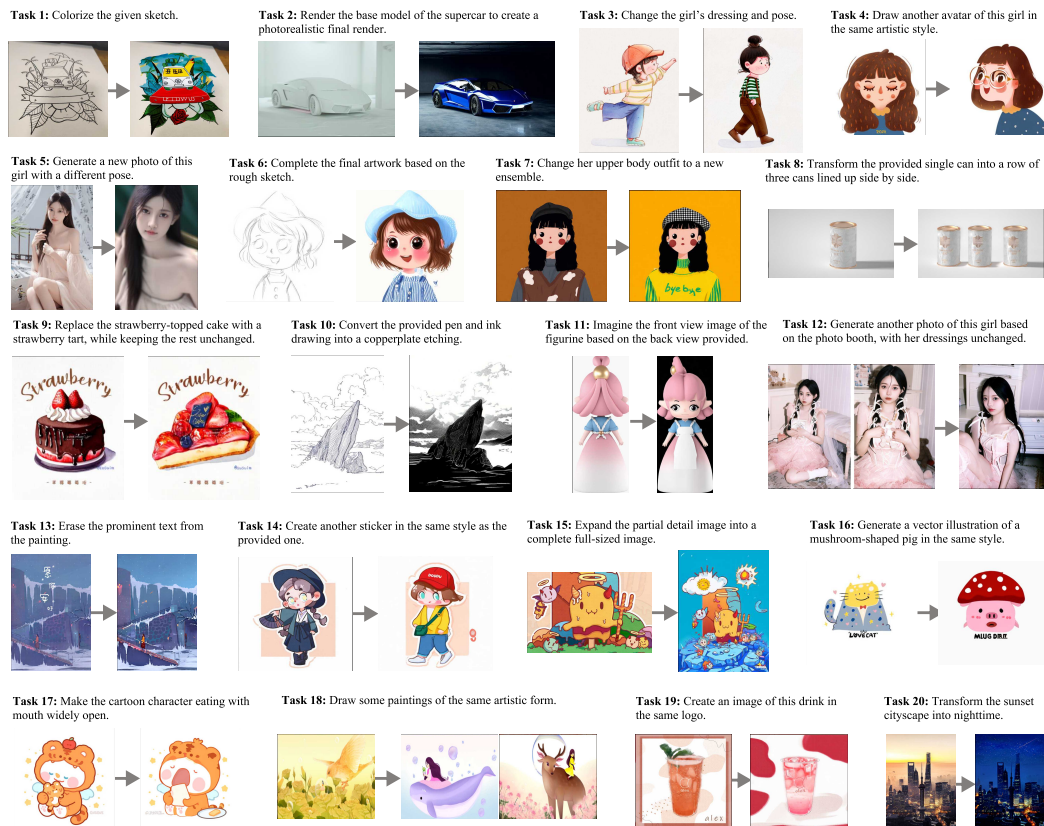


Figure 2: When conditioned on a subset of the group data, Group Diffusion Transformers could perform conditional group generation in the inpainting fashion. Note that the model is automatically generalized to these tasks after unsupervised training on image groups. Textual descriptions of images are 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 $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, where each element is conditioned on its respective context (*e.g.*, *image descriptions*) $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n\}$. 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.

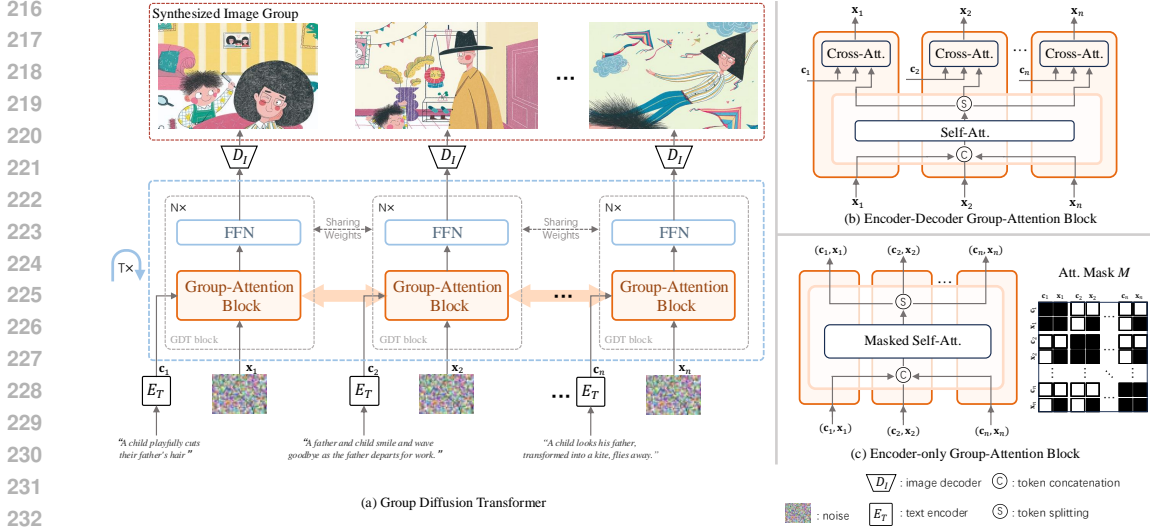


Figure 3: **The overview of Group Diffusion Transformer, which takes minimal adaptations for the encoder-decoder and encoder-only visual generation architectures.** We make a straightforward modification on self-attention blocks by concatenating image tokens across group inputs, allowing to learn inter-image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

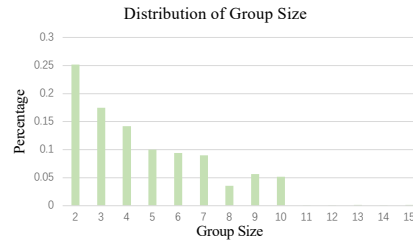


Figure 4: **Distribution of group size in our training dataset.**

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.

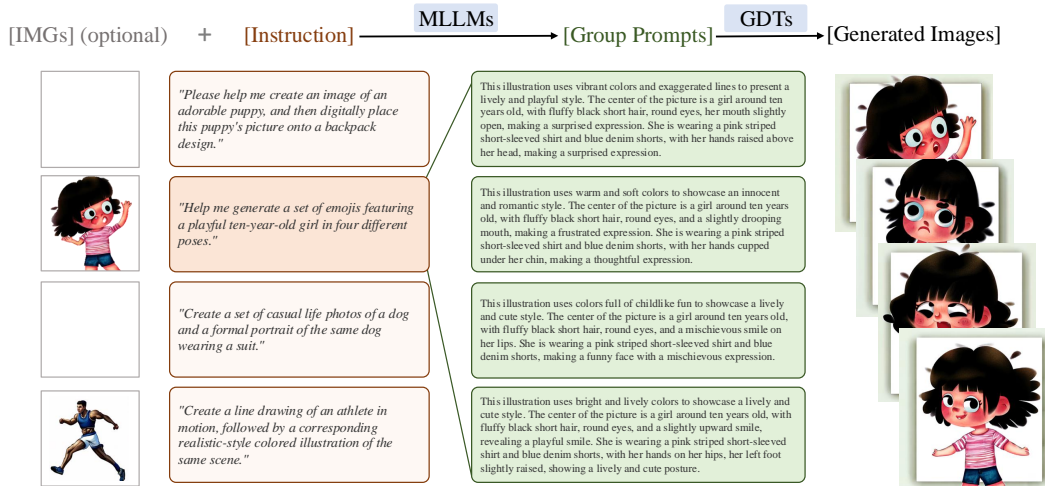


Figure 7: We build a user interface that automatically converts the user instruction into group prompts using MLLMs, which is useful in the inference stage of GDTs.

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

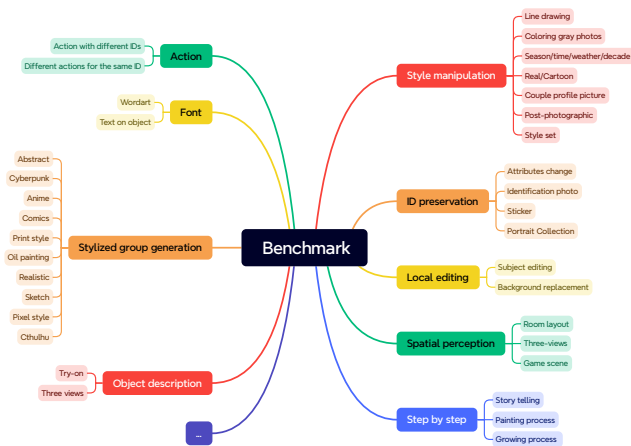


Figure 6: Overview of our benchmark, covering about 30 distinct types of generation tasks.

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.

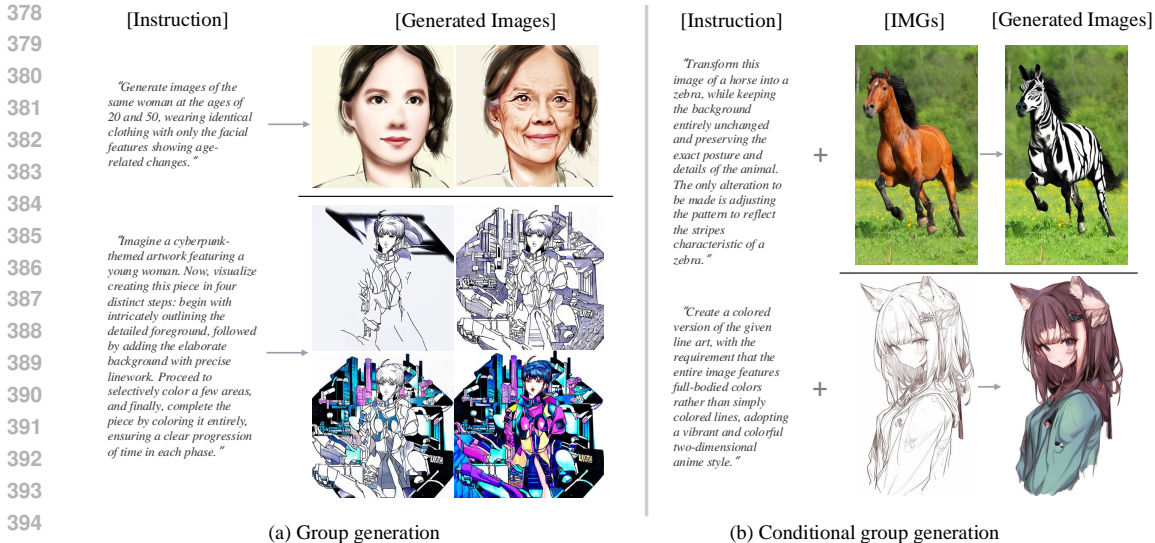


Figure 8: Generated results of GDTs on our benchmark, including group generation and conditional group generation.

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 SEDIT 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 Vilevski 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.

Moreover, 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

- 540
541
542 Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng
543 Shen, Mohamed Awadalla, Silvio Savarese, et al. Mint-1t: Scaling open-source multimodal data
544 by 10x: A multimodal dataset with one trillion tokens. *arXiv preprint arXiv:2406.11271*, 2024.
- 545 Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas
546 Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023. URL <https://arxiv.org/abs/2302.07121>.
- 547
548 Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
549
- 550 Hernan Carrillo, Michaël Clément, Aurélie Bugeau, and Edgar Simo-Serra. Diffusart: Enhancing
551 line art colorization with conditional diffusion models. In *Proceedings of the IEEE/CVF Confer-*
552 *ence on Computer Vision and Pattern Recognition*, pp. 3486–3490, 2023.
- 553 Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite:
554 Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on*
555 *Graphics*, 42:1–10, 07 2023.
556
- 557 Changyou Chen, Han Ding, Bunyamin Sisman, Yi Xu, Ouye Xie, Benjamin Z. Yao, Son Dinh
558 Tran, and Belinda Zeng. Diffusion models for multi-task generative modeling. In *The Twelfth*
559 *International Conference on Learning Representations*, 2024a. URL <https://openreview.net/forum?id=cbv0sBIZh9>.
560
- 561 Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang,
562 James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion
563 transformer for photorealistic text-to-image synthesis. *ArXiv*, abs/2310.00426, 2023a. URL
564 <https://api.semanticscholar.org/CorpusID:263334265>.
- 565 Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao
566 Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image
567 diffusion models, 2023b.
568
- 569 Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, and Zhendong Mao. Dreami-
570 dentity: Enhanced editability for efficient face-identity preserved image generation. *Proceedings*
571 *of the AAAI Conference on Artificial Intelligence*, 38(2):1281–1289, Mar. 2024b.
- 572 Siying Cui, Jia Guo, Xiang An, Jiankang Deng, Yongle Zhao, Xinyu Wei, and Ziyong Feng.
573 Idadapter: Learning mixed features for tuning-free personalization of text-to-image models, 2024.
574 URL <https://arxiv.org/abs/2403.13535>.
- 575 Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon
576 Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Rade-
577 novic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran
578 Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Va-
579 jda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a
580 haystack, 2023. URL <https://arxiv.org/abs/2309.15807>.
- 581 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
582 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
583 *arXiv preprint arXiv:2407.21783*, 2024.
584
- 585 Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi,
586 Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle
587 Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers
588 for high-resolution image synthesis. *ArXiv*, abs/2403.03206, 2024a. URL <https://api.semanticscholar.org/CorpusID:268247980>.
589
- 590 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam
591 Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion En-
592 glish, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow
593 transformers for high-resolution image synthesis, 2024b. URL <https://arxiv.org/abs/2403.03206>.

- 594 Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel
595 Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual
596 inversion, 2022. URL <https://arxiv.org/abs/2208.01618>.
- 597 Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint*
598 *arXiv:2205.15463*, 2022.
- 600 Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhang, and Tobias
601 Hinz. Personalized residuals for concept-driven text-to-image generation. In *Proceedings of the*
602 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8186–8195,
603 June 2024.
- 604 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In
605 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Infor-*
606 *mation Processing Systems*, volume 33, pp. 6840–6851, 2020.
- 607 Ming-Yang Ho, Che-Ming Wu, Min-Sheng Wu, and Yufeng Jane Tseng. Every pixel has its mo-
608 ments: Ultra-high-resolution unpaired image-to-image translation via dense normalization. *arXiv*
609 *preprint arXiv:2407.04245*, 2024.
- 611 Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-
612 Jen Cham. Cocktail: Mixing multi-modality controls for text-conditional image generation. *arXiv*,
613 2023.
- 614 Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative
615 and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*,
616 2023.
- 617 Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Chang-
618 sheng Xu. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Trans-*
619 *actions on Neural Networks and Learning Systems*, 2024.
- 620 Ze Jin and Zorina Song. Generating coherent comic with rich story using chatgpt and stable diffu-
621 sion. *arXiv preprint arXiv:2305.11067*, 2023.
- 622 Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-
623 to-image models with a single image pair. *arXiv preprint arXiv:2405.01536*, 2024.
- 624 Ziyi Kou, Shichao Pei, Yijun Tian, and Xiangliang Zhang. Character as pixels: A controllable
625 prompt adversarial attacking framework for black-box text guided image generation models. In
626 *IJCAI*, pp. 983–990, 2023.
- 627 Manoj Kumar, Neil Houlsby, and Emiel Hoogeboom. Semantica: An adaptable image-conditioned
628 diffusion model. *arXiv preprint arXiv:2405.14857*, 2024.
- 629 Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept
630 customization of text-to-image diffusion. 2023.
- 631 Black Forest Labs. Flux.1, 2024. URL [https://github.com/black-forest-labs/](https://github.com/black-forest-labs/flux)
632 `flux`.
- 633 Dongxu Li, Junnan Li, and Steven Hoi. BLIP-diffusion: Pre-trained subject representation for
634 controllable text-to-image generation and editing. In *Thirty-seventh Conference on Neural*
635 *Information Processing Systems*, 2023a. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=g6We1SwaY9)
636 `g6We1SwaY9`.
- 637 Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li,
638 and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the*
639 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521,
640 June 2023b.
- 641 Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Pho-
642 tomaker: Customizing realistic human photos via stacked id embedding. In *IEEE Conference on*
643 *Computer Vision and Pattern Recognition (CVPR)*, 2024a.

- 648 Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng Deng, Yingfang Zhang,
649 Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen
650 Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li,
651 Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu
652 Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen,
653 Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and
654 Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained
655 chinese understanding, 2024b. URL <https://arxiv.org/abs/2405.08748>.
- 656 Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color:
657 Multimodal diffusion-based interactive image colorization. *arXiv preprint arXiv:2402.10855*,
658 2024.
- 659 Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Xintao Wang, Yujiu Yang, and
660 Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv*
661 *preprint arXiv:2312.00330*, 2023a.
- 662 Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang.
663 Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint*
664 *arXiv:2309.03453*, 2023b.
- 665 Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren
666 Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized gener-
667 ation. In *International Conference on Machine Learning*, 2023c. URL [https://api.](https://api.semanticscholar.org/CorpusID:257427549)
668 [semanticscholar.org/CorpusID:257427549](https://api.semanticscholar.org/CorpusID:257427549).
- 669 Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent
670 diffusion for pose-guided person image synthesis. In *Proceedings of the IEEE/CVF Conference*
671 *on Computer Vision and Pattern Recognition*, pp. 6420–6429, 2024.
- 672 Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo
673 Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module.
674 *arXiv preprint arXiv:2311.05556*, 2023.
- 675 Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon.
676 Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International*
677 *Conference on Learning Representations*, 2021. URL [https://api.semanticscholar.](https://api.semanticscholar.org/CorpusID:245704504)
678 [org/CorpusID:245704504](https://api.semanticscholar.org/CorpusID:245704504).
- 679 Paritosh Mittal, Kunal Aggarwal, Pragma Paramita Sahu, Vishal Vatsalya, Soumyajit Mitra, Vikrant
680 Singh, Viswanath Veera, and Shankar M Venkatesan. Photo-realistic emoticon generation us-
681 ing multi-modal input. In *Proceedings of the 25th International Conference on Intelligent User*
682 *Interfaces*, pp. 254–258, 2020.
- 683 Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan.
684 T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion
685 models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–
686 4304, 2024.
- 687 Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew,
688 Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with
689 text-guided diffusion models, 2022.
- 690 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
691 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-
692 low instructions with human feedback. *Advances in neural information processing systems*, 35:
693 27730–27744, 2022.
- 694 William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of*
695 *the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October
696 2023.

- 702 Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext:
703 Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*,
704 2024.
- 705 Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Car-
706 los Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for
707 controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.
- 708 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language
709 models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 710 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi
711 Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text
712 transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- 713 Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-
714 conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022a. URL <https://api.semanticscholar.org/CorpusID:248097655>.
- 715 Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-
716 conditional image generation with clip latents, 2022b. URL <https://arxiv.org/abs/2204.06125>.
- 717 Benjamin Rodatz, Ian Fan, Tuomas Laakkonen, Neil John Ortega, Thomas Hoffman, and Vin-
718 cent Wang-Mascianica. A pattern language for machine learning tasks. *arXiv preprint*
719 *arXiv:2407.02424*, 2024.
- 720 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-
721 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Con-*
722 *ference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.
- 723 Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman.
724 Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.
- 725 Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kam-
726 yar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Sal-
727 imans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffu-
728 sion models with deep language understanding, 2022.
- 729 Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. ViPer: Visual
730 personalization of generative models via individual preference learning. *ECCV*, 2024.
- 731 Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image
732 synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*, 2023.
- 733 Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen,
734 Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base
735 model. *arXiv preprint arXiv:2310.15110*, 2023.
- 736 James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia
737 Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv*
738 *preprint arXiv:2304.06027*, 2023.
- 739 Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred
740 Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any
741 style. *arXiv preprint arXiv:2306.00983*, 2023.
- 742 Yiren Song, Shijie Huang, Chen Yao, Xiaojun Ye, Hai Ci, Jiaming Liu, Yuxuan Zhang, and
743 Mike Zheng Shou. Processpainter: Learn painting process from sequence data, 2024. URL
744 <https://arxiv.org/abs/2406.06062>.
- 745 Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang,
746 Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models
747 are in-context learners, 2024.

- 756 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
757 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and
758 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- 759
760 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-
761 lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open founda-
762 tion and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- 763 Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously condi-
764 tioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, 2023.
- 765
766 Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion mod-
767 els. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.
- 768
769 Chi Wang, Min Zhou, Tiezheng Ge, Yuning Jiang, Hujun Bao, and Weiwei Xu. Cf-font: Content
770 fusion for few-shot font generation. In *Proceedings of the IEEE/CVF Conference on Computer
771 Vision and Pattern Recognition*, pp. 1858–1867, 2023a.
- 772
773 Dejiang Wang, Zhuoran Zhai, Ngai Cheong, and Li Peng. Script-generated picture book technology
774 based on large language models and aigc. In *Proceedings of the 7th International Conference on
775 Digital Technology in Education*, pp. 104–108, 2023b.
- 776
777 Qiang Wang, Di Kong, Fengyin Lin, and Yonggang Qi. Diffsketching: Sketch control image syn-
778 thesis with diffusion models. *arXiv preprint arXiv:2305.18812*, 2023c.
- 779
780 Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-
781 preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.
- 782
783 Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding
784 visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings
785 of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.
- 786
787 Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer:
788 Tuning-free multi-subject image generation with localized attention. *arXiv*, 2023.
- 789
790 Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape
791 guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on
792 Computer Vision and Pattern Recognition*, pp. 22428–22437, 2023.
- 793
794 Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion:
795 Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF Interna-
796 tional Conference on Computer Vision (ICCV)*, pp. 7754–7765, October 2023.
- 797
798 Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free
799 diffusion: Taking” text” out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF
800 Conference on Computer Vision and Pattern Recognition*, pp. 8682–8692, 2024.
- 801
802 Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable
803 diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint
804 arXiv:2308.14469*, 2023.
- 805
806 Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. Fontdiffuser:
807 One-shot font generation via denoising diffusion with multi-scale content aggregation and style
808 contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2024.
- 809
810 Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt
811 adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023a.
- 812
813 Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt
814 adapter for text-to-image diffusion models. 2023b.
- 815
816 Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free
817 energy-guided conditional diffusion model. *Proceedings of the IEEE/CVF International Confer-
818 ence on Computer Vision (ICCV)*, 2023.

- 810 Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image
811 colorization with text guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11,
812 2023.
- 813 Jan Zdenek and Hideki Nakayama. Handwritten text generation with character-specific encoding for
814 style imitation. In *International Conference on Document Analysis and Recognition*, pp. 313–329.
815 Springer, 2023.
- 816 Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image
817 diffusion models, 2023.
- 818 Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christo-
819 pher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer
820 language models. *arXiv preprint arXiv:2205.01068*, 2022.
- 821 Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-
822 Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in
823 Neural Information Processing Systems*, 36, 2024.
- 824
825
826 Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Con-
827 sistent self-attention for long-range image and video generation. 2024.
- 828
829

830 A APPENDIX

831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917



The image presents a large three-dimensional letter 'J', made of polished metal with a smooth surface and a strong mirror reflection. The interior of the letter reflects a blurry cityscape or mountain range, primarily in dark blue and gray-white tones. The letter is placed against a rough gray cement wall, creating a sharp contrast that enhances the metallic texture and three-dimensionality.

The image is dominated by a large three-dimensional letter 'I', made of polished metal with a mirror effect, reflecting the surrounding light. Inside the letter is a clear black and white silhouette of a city, with densely packed buildings. The bottom is dark blue, contrasting with the bright white above. The background is a rough gray cement wall, creating visual tension with the metallic letter.

The image shows a large three-dimensional letter 'Y', also made of polished metal with a mirror reflection. The letter reflects a relatively clear city outline; the buildings are arranged irregularly, and the overall color is dark. The bottom of the letter transitions to dark blue, contrasting with the gray cement wall in the background, creating an industrial and modern feel. The letter is three-dimensional and has a strong texture.

The main subject of the image is a transparent orange letter 'B', with a fine water droplet-like texture on its surface, giving it a crystal-clear texture. The letter 'B' is located in the center of the image, surrounded by bright orange rays of light radiating outwards, creating a dazzling light effect. Small orange droplets are scattered in the background, echoing the texture on the surface of the letter. The overall tone is warm, bright, and full of vitality.

The main subject of the image is a transparent orange letter 'O', with a fine water droplet-like texture on its surface, giving it a crystal-clear texture. The letter 'O' is located in the center of the image, surrounded by bright orange rays of light radiating outwards, creating a dazzling light effect. Small orange droplets are scattered in the background, echoing the texture on the surface of the letter. The overall tone is warm, bright, and full of vitality.



A young woman with long red hair is shown with her eyes closed, wearing a beige sweater with a white collar underneath and round glasses. The background features a sunset and a purple backdrop.

A young woman with long red hair is shown with her eyes closed, wearing a beige sweater with a white collar underneath and round glasses. The background features a sunset and a purple backdrop, and her hair is blowing in the wind.

A young woman with long red hair is shown wearing a beige sweater with a white collar underneath and round glasses. She turns her head to look at the camera, with a sunset and a purple backdrop in the background.



In the photo, a young couple is all smiles. The girl is wearing a white slip dress and the boy is wearing a black suit with a white shirt and black tie. They are holding hands and look very sweet and happy.

In the photo, a young couple is in an intimate pose. The girl is wearing a white slip dress and the boy is wearing a black suit. They are looking into each other's eyes and the girl has a sweet smile on her face, overflowing with happiness.

In the photo, a young couple is in an intimate pose. The girl is wearing a white slip dress with her back to the camera, revealing her bare back. The boy is wearing a black suit and looking at the girl's back affectionately.



A young woman floats in the water, her head slightly lifted and her gaze directed upwards. Her dark hair spreads around her, contrasting sharply with her fair skin and vibrant red lipstick. The water surrounds her face and hair, creating a dreamlike atmosphere.

A young woman floats in the water with her eyes closed. Her dark hair spreads around her, contrasting with her fair skin and vibrant red lipstick. The water surrounds her face, creating a peaceful and serene atmosphere.

A young woman leans on a chest, gazing directly at the viewer. She has fair skin, her brown hair falls loosely around her shoulders, and her dark eyes hold a gentle expression. She wears a light-colored shirt, the neckline slightly open, revealing the graceful lines of her neck.

A young woman stands by the window, her head tilted slightly upwards as if feeling the caress of a gentle breeze. She has fair skin, her brown hair falls naturally, and her dark eyes are filled with hope for the future. She wears a light-colored shirt with loose, comfortable sleeves and dark pants, creating a simple and elegant look.



A fashion photography piece showcasing a female model in a pink printed maxi dress layered over with a long white furry coat. The model has long brown hair and sophisticated makeup; she stands against a light-colored backdrop in an elegant pose with her hands in her pockets. The dress has a V-neck design, is lightweight in texture, and features a soft print. The coat is fluffy and soft, with a high-end feel. The model is wearing gold strapped heels.

Fashion photography showcasing a female model in a champagne-colored strapless A-line dress. The dress is smooth in texture, with a strong drape, and an A-line skirt, cinched at the waist with a golden belt. The model's hairstyle and makeup are consistent with the previous image, and the background is similarly simple and light-colored. The model's pose is elegant, with her gaze directed forward. She wears gold heels that complement the dress.

A fashion photography image displaying a female model in a burgundy velvet maxi dress. The dress is a halter neck style, with a defined waist, and an asymmetric hemline, trailing on one side while revealing the ankle on the other. The dress is thick in texture, with a rich color, and the skirt flows slightly. The model's hairstyle and makeup remain consistent, against a simple light-colored background. The model's posture is elegant, and she wears gold high-heeled shoes.

A fashion photography piece featuring a female model in a purplish-red velvet slip dress. The dress is simply cut and form-fitting, accented with a thin silver belt at the waist. The model maintains the consistent hairstyle and makeup from other images, against a light-colored background. The model poses elegantly, with a confident smile. Her makeup is refined, with simple earrings. The model wears light gold heeled shoes.

Figure 1: Detailed results of Group Diffusion Transformers.

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971



A futuristic cyberpunk car faces the camera, showcasing its front design. The car boasts sleek lines, intricate details, and dazzling lighting effects. The backdrop is a technologically advanced cityscape at night.

A futuristic cyberpunk car is presented sideways, showcasing its side profile and streamlined body. The car boasts sleek lines, intricate details, and dazzling lighting effects. The backdrop is a technologically advanced cityscape at night.

A futuristic cyberpunk car is seen from above, showcasing its roof design and overall layout. The car boasts sleek lines, intricate details, and dazzling lighting effects. The backdrop is a technologically advanced cityscape at night.

The image shows a modern minimalist living room from an overhead perspective. A beige fabric L-shaped sofa occupies the center of the space, complemented by a dark wood coffee table and a light-colored rug. Decorative items are placed on the coffee table. A floor-to-ceiling window dominates one wall, showcasing lush greenery outside. An abstract artwork hangs on the wall, and a wooden side table is visible in a corner. The overall color scheme is soft, creating a relaxed and comfortable atmosphere.

The image depicts a modern minimalist living room. A beige fabric sofa is centrally positioned, featuring smooth lines and soft, comfortable material. A dark wood coffee table with a simple and elegant design sits in front of the sofa. A floor-to-ceiling window occupies one wall, revealing a pleasant outdoor view. An abstract artwork hangs on the wall, next to a wooden side cabinet displaying a blue vase. The overall space is spacious and bright, with a clean and sophisticated design style.



A front view of a Minecraft-style game scene, the picture is full of pixelated style. In the foreground is green grass dotted with colorful flowers. Not far away is a house built of oak, with a path paved with stone bricks in front of the house. Around the house are planted several lush trees with leaves in different shades of green. Several white clouds float in the sky.

A side view of a Minecraft-style game scene, the picture is full of pixelated style. In the foreground is green grass dotted with colorful flowers. The grass stretches into the distance, where it meets the blue sky. On the left side of the picture, you can see a house built of oak, with a path paved with stone bricks in front of the house. Several lush trees are planted next to the house, with leaves in different shades of green. Several white clouds float in the sky.

A top view of a Minecraft-style game scene, the picture is full of pixelated style. In the center of the picture is a green grass with colorful flowers dotted on it. There is a house built of oak on the grass, with a path paved with stone bricks in front of the house. Several lush trees planted around are the house, with leaves in different shades of green. A winding river can be seen beside the house. Several white clouds float in the sky.

A fluffy squirrel, dressed in a brown plaid shirt and green pants, stands on the grass. It has a wide grin, wide-open eyes, and holds a giant pine cone, looking delighted.

A fluffy squirrel, dressed in blue overalls, lies on top of a giant pine cone. It has a frown, wide-open eyes, and a slightly open mouth, looking angry.

A fluffy squirrel, dressed in a brown plaid shirt and green pants, stands on the grass. It has wide-open eyes and a slightly open mouth, looking surprised.

A fluffy squirrel, dressed in a brown plaid shirt and green pants, sits on a branch. It has a wide grin, wide-open eyes, and looks happy. A small bird sits on the branch next to it, and another bird sits on a higher branch.



The picture depicts a sudden downpour in the countryside. Large raindrops fall from the sky as people on the hillside diligently harvesting the ripe wheat. They tie the wheat stalks into bundles and stack them at the edge of the field. In the distance are rolling mountains, with houses and farmstead at the foot of the mountains, forming a beautiful rural landscape.

The picture depicts a busy harvest scene in the countryside. In the golden wheat fields, farmers are diligently harvesting the ripe wheat. They tie the wheat stalks into bundles and stack them at the edge of the field. In the distance are rolling mountains, with houses and farmstead at the foot of the mountains, forming a beautiful rural landscape.

The picture depicts a tranquil rural scene. A country road winds through the picture, flanked by several houses surrounded by lush greenery. On a table by the road lie slices of watermelon with a few birds pecking at the crumbs on the ground. A dog lies under the table, seemingly anticipating the delicious treat.

The picture depicts a bountiful harvest scene in the countryside. Golden wheat fields stretch as far as the eye can see, with rolling mountains and a few houses in the distance. A scarecrow stands in the wheat field, guarding the upcoming harvest. A few birds perch on a fence by the field, seemingly admiring the beautiful scenery.

A blue-green car is driving on the road with two rabbits in it.

A red car is driving on the road with two mice in it.

A blue car is driving on the road with four animals inside: a giraffe, a tiger, a frog, and a teddy bear.

Figure 2: Detailed results of Group Diffusion Transformers.

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025



Figure 3: Detailed results of Group Diffusion Transformers.

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079



A cartoon character, resembling Buddha, wearing a golden lotus headdress and pink monk robes. His hands are hidden in his sleeves. He has a slight smile and closed eyes, showing contentment.

A cartoon character, resembling Buddha, wearing a golden lotus headdress and pink monk robes. His hands are hidden in his sleeves. He has a furrowed brow and downturned mouth, showing sadness.

A cartoon character, resembling Buddha, wearing a golden lotus headdress and pink monk robes. His hands are hidden in his sleeves. He has a big smile and closed eyes, showing joy.

A cartoon character, resembling Buddha, wearing a golden lotus headdress and pink monk robes. His hands are hidden in his sleeves. His eyes are closed and his eyebrows are slightly furrowed, showing exhaustion.

A photo of a square glass bottle containing a light brown liquid. The bottle is sealed with a wooden stopper and has a label with Chinese characters and design elements. The bottle is surrounded by green plants, creating a natural feel.

A photo of a square glass bottle containing a light brown liquid. The bottle is sealed with a wooden stopper and has a label with Chinese characters and design elements. It's set against a dark background and surrounded by scattered orange petals, creating a professional studio ambiance.

A product photography image with a warm orange background. The main focus is a golden gift box with a fan-shaped cutout pattern and landscape painting decorations; three golden mooncakes are visible inside. Next to the gift box is a red tray, also containing three mooncakes. Abstract white mountain-shaped decorations are on the right. On the left is a red rectangular object and a red round tray with a mooncake on it. The overall style of the image is simple and clear, highlighting the elegance and sophistication of the gift box. The color scheme is harmonious, creating a comfortable visual experience.

A product photography image with a warm orange background. The center of the image features a white gift box with a fan-shaped cutout pattern and landscape painting decorations; three golden mooncakes are visible inside. Abstract white mountain-shaped decorations are on the right. On the left is a red rectangular object and a red round tray with a mooncake on it. The overall style of the image is simple and clear, highlighting the elegance and sophistication of the gift box. The color scheme is harmonious, creating a comfortable visual experience.

Figure 4: Detailed results of Group Diffusion Transformers.

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

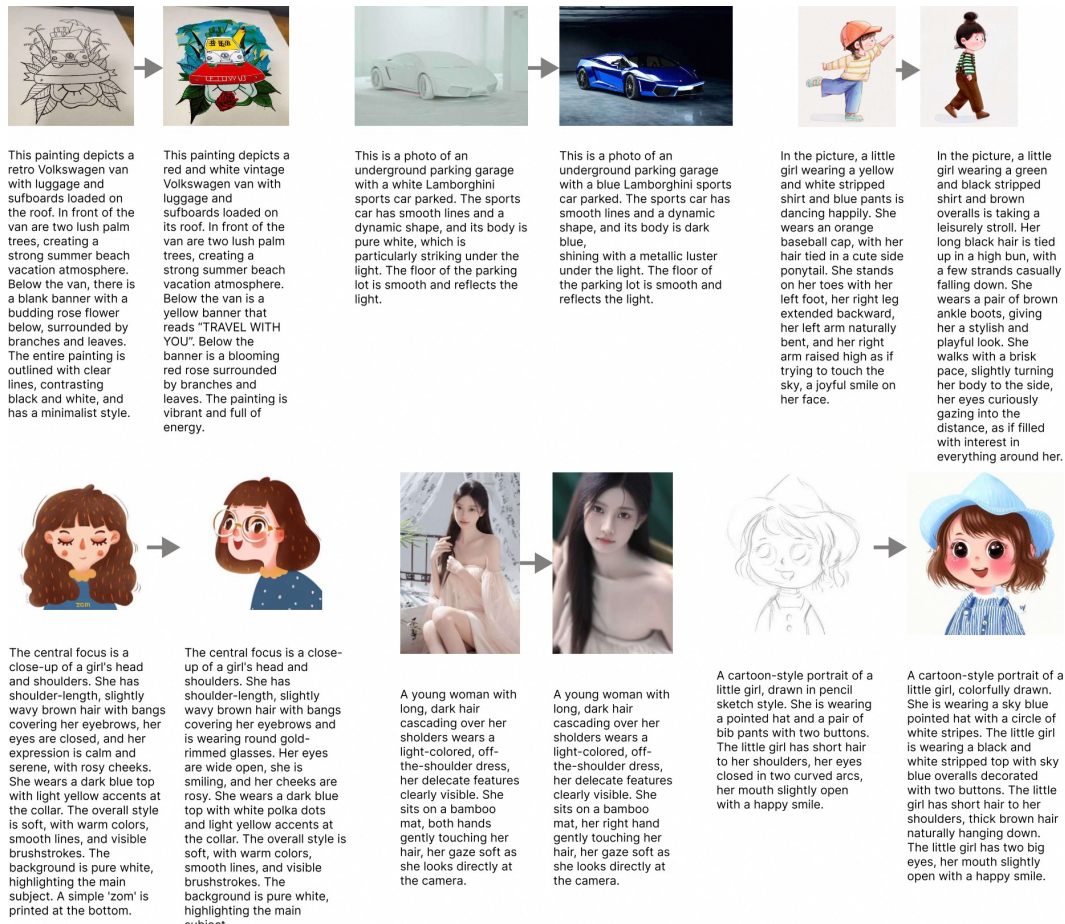


Figure 5: Detailed results of Group Diffusion Transformers.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

<p>A girl with black bob hair and bangs is wearing a black beret. She has large eyes, a round face, rosy cheeks, and is wearing silver earrings. She is wearing a brown vest with white cloud patterns over a blue turtleneck.</p>	<p>A girl with black bob hair and bangs is wearing a black checkered beret. She has large eyes, a round face, rosy cheeks, and is wearing gold earrings. She is wearing a yellow sweater with the words 'byebye' printed on it and green striped collar and cuffs.</p>	<p>A product display image with a pure white background. A cylindrical metal can is placed in the center. The can is white with a delicate textured pattern on its surface, resembling clouds or petals, with a golden metallic sheen. The lid is made of rose gold metal. The center of the can features a golden brand logo, including the brand name 'PHYSICAL ART' and the Chinese brand name '悠然'. The bottom of the can is marked with 'CONCERT 45ML'. The overall style is simple, elegant, and refined.</p>	<p>A product display image with a pure white background. Three identical cylindrical metal cans are arranged side by side. The cans are white with a delicate textured pattern on their surface, resembling clouds or petals, with a golden metallic sheen. The lids are made of rose gold metal. The center of each can features a golden brand logo, including the brand name 'PHYSICAL ART' and the Chinese brand name '悠然'. The bottom of each can is marked with 'CONCERT 45ML'. The overall style is simple, elegant, and refined.</p>	<p>The center of the image is a chocolate cake. The bottom of the cake is red, the top is covered with thick chocolate sauce, and it's decorated with fluffy cream and two fresh strawberries. The cake rests on a dark base. Above the image is the title 'Strawberry', with an elegant and flowing font. The background is simple, the overall style is fresh and sweet, creating a comfortable visual experience. The watercolor painting technique gives the image a soft color transition and a light texture.</p>	<p>The main subject of the image is a slice of strawberry tart. The tart crust is topped with bright red strawberries, decorated with a few blueberries and cherries. There's a golden chocolate decoration on the strawberries. The cross-section of the tart shows a light layering and the texture of the filling. Above the image is the title 'Strawberry', consistent with the first image. The background is equally simple, and the watercolor painting technique creates a light and dreamy atmosphere.</p>
<p>This illustration depicts a huge rock standing on the coast in a pen-and-ink sketch style. The surface of the rock is rough, with rich texture details, and the use of dense lines to depict the effect of light and shadow. Behind the rock is the rough sea, and the rolling mountains can be vaguely seen in the distance. The sky is covered with clouds of various shapes.</p>	<p>This illustration depicts a huge rock standing on the coast in the style of an engraving. The surface of the rock is rough, with rich texture details, and the use of dense lines to depict the effect of light and shadow. Behind the rock is the rough sea, and the rolling mountains can be vaguely seen in the distance. The sky is filled with clouds of various shapes, with sunlight shining through the clouds.</p>	<p>This is a 3D rendered image showcasing the back view of a cartoon girl. She has pink hair tied up in a high bun, adorned with a golden spherical ornament. She is wearing a blue top with a white apron tied around her waist, featuring rope-like detailing. Her dress is pink, gradually fading to white at the bottom. She has pointed ears, exhibiting an overall cute and sweet style.</p>	<p>This is a 3D rendered image showcasing the front view of a cartoon girl. She has pink hair tied up in a high bun, adorned with a golden spherical ornament. She has large eyes and cute pointed ears. She is wearing a blue short-sleeved top with a white apron. She is wearing a pink dress, gradually fading to white at the bottom. The overall style of the girl is cute and sweet.</p>	<p>A young Asian woman with long, black hair sits on the floor wearing a pink lace dress with white ribbon decorations tied in bows on her head. She holds a storybook and looks at the viewer with clear eyes. The room features white wood paneling and a white cabinet.</p>	<p>A young Asian woman with long, black hair sits on the floor wearing a pink lace dress with white ribbon decorations tied in bows on her head. She rests one hand on her leg and looks directly at the viewer with clear eyes. The room features white wood paneling and a white cabinet.</p>
<p>The scene presents a tranquil winter nightscape. Snow-laden stone buildings stand tall on either side, with dense icicles hanging from their roofs. A woman in a reddish-brown coat ascends a stone staircase, carrying a warm lantern. Dim yellowish lights emanate from the buildings, contrasting with the twinkling stars in the sky. Fine snowflakes fall from the sky, and a thin layer of snow covers the ground. The overall color palette is cool, creating a serene and peaceful atmosphere. In the distance, another figure can be vaguely seen standing at a building entrance. A lone, bare tree is visible in the lower right corner. Vertical Chinese characters reading '愿你安好' are displayed at the top.</p>	<p>The image depicts a cold winter night scene. Tall buildings on both sides are covered with thick snow and long icicles. A stone staircase winds upwards, with a figure in an orange-yellow coat walking along it. The buildings' lights are warm and yellowish, contrasting with the sparse stars in the night sky. Tiny snowflakes fall from the sky, and the ground is covered in snow. The color palette is cool, but the lights provide a touch of warmth, creating a quiet and slightly mysterious atmosphere. A bare tree and a partial view of the building interiors are visible in the background. The overall style is dreamy and slightly impressionistic.</p>	<p>The image depicts a cold winter night scene. Tall buildings on both sides are covered with thick snow and long icicles. A stone staircase winds upwards, with a figure in an orange-yellow coat walking along it. The buildings' lights are warm and yellowish, contrasting with the sparse stars in the night sky. Tiny snowflakes fall from the sky, and the ground is covered in snow. The color palette is cool, but the lights provide a touch of warmth, creating a quiet and slightly mysterious atmosphere. A bare tree and a partial view of the building interiors are visible in the background. The overall style is dreamy and slightly impressionistic.</p>	<p>The image presents an adorable chibi girl dressed in dark blue clothing. She wears a dark blue hat, a dark blue jacket, a white top underneath, a dark blue skirt, and black ankle boots. She has two braided pigtails and holds a dark purple folding fan. Her eyes are clear and bright, and her overall style is fashionable with traditional elements. The background is light orange with small dots around the edges, creating a fresh and cute style.</p>	<p>The image shows an adorable chibi girl in casual attire. She sports a red baseball cap with 'DOUDU' printed on it, a yellow jacket over a white top, blue jeans, and white sneakers, along with a dark green crossbody bag. She has shoulder-length brown hair, slightly rosy cheeks, and a somewhat shy expression, creating a youthful and lively overall style. The background is light orange with small dots around the edges, maintaining a fresh and cute style.</p>	<p>The image shows an adorable chibi girl in casual attire. She sports a red baseball cap with 'DOUDU' printed on it, a yellow jacket over a white top, blue jeans, and white sneakers, along with a dark green crossbody bag. She has shoulder-length brown hair, slightly rosy cheeks, and a somewhat shy expression, creating a youthful and lively overall style. The background is light orange with small dots around the edges, maintaining a fresh and cute style.</p>

Figure 6: Detailed results of Group Diffusion Transformers.

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

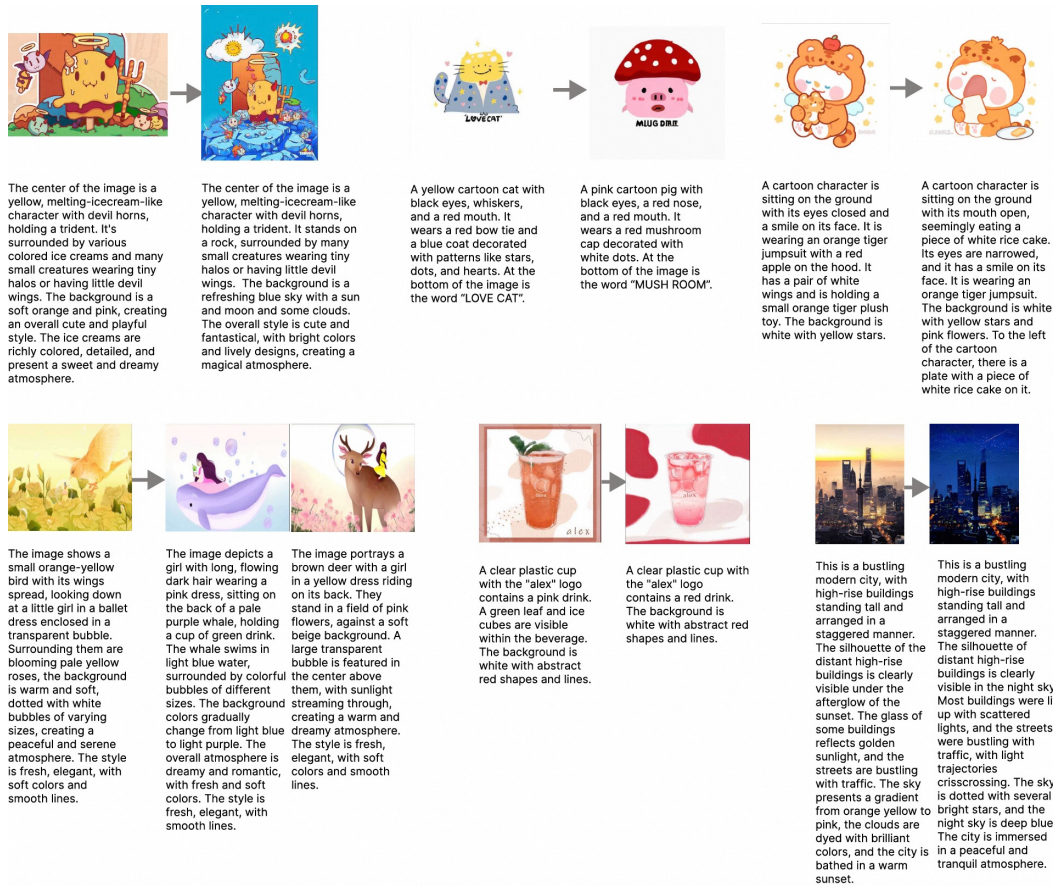


Figure 7: Detailed results of Group Diffusion Transformers.