Ranni: Taming Text-to-Image Diffusion for Accurate Instruction Following

Yutong Feng[1], Biao Gong[1], Di Chen[1], Yujun Shen[2], Yu Liu[1], Jingren Zhou[1]
[1]Alibaba Group [2]Ant Group
{fengyutong.fyt, gongbiao.gb, guangpan.cd, ly103369, jingren.zhou}@alibaba-inc.com
shenyujun0302@gmail.com

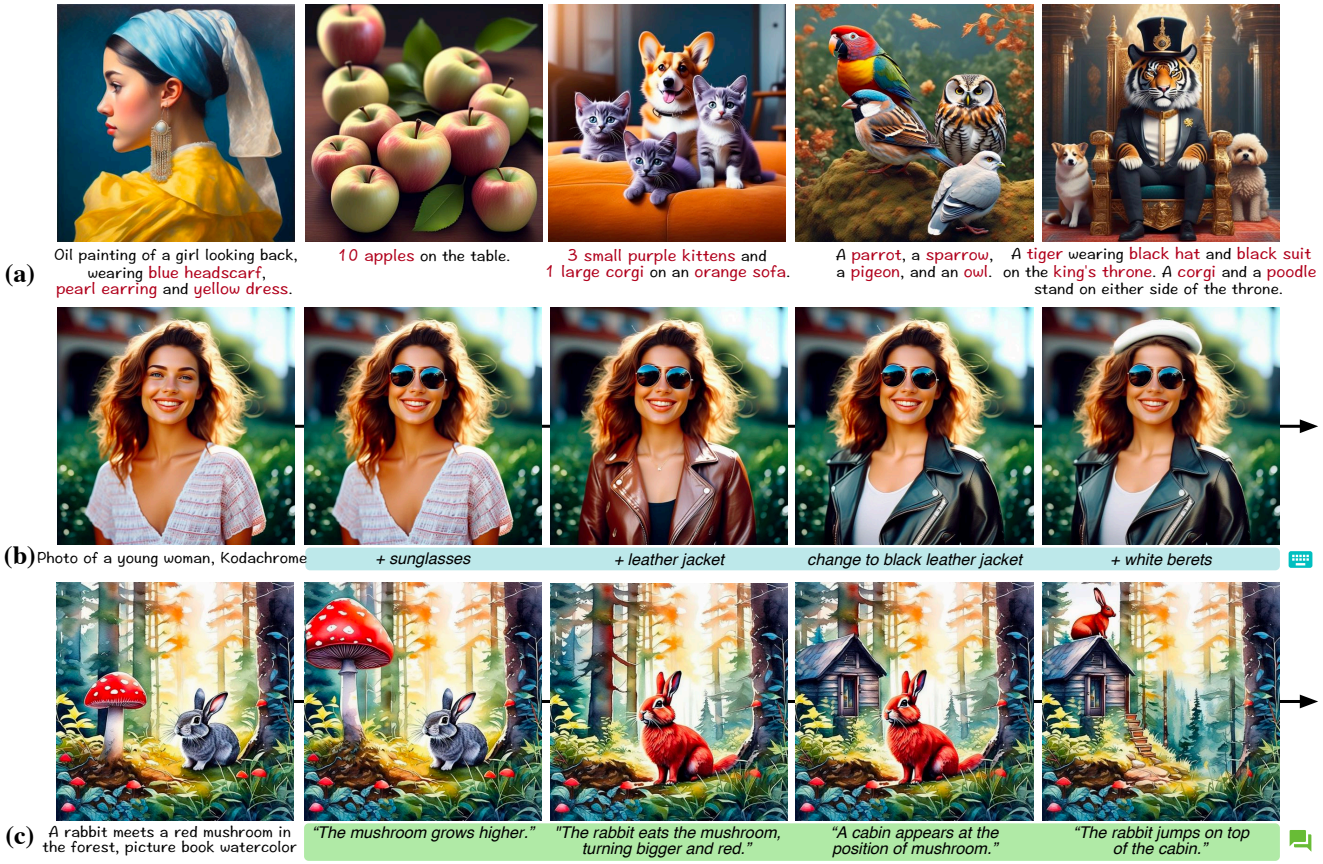


Figure 1. Samples generated by `Ranni` with different interaction manners, including **(a) direct generation** with accurate prompt following, **(b) continuous generation** with progressive refinement, and **(c) chatting-based generation** with text instructions.

Abstract

Existing text-to-image (T2I) diffusion models usually struggle in interpreting complex prompts, especially those with quantity, object-attribute binding, and multi-subject descriptions. In this work, we introduce a ***semantic panel*** *as the middleware in decoding texts to images, supporting the generator to better follow instructions. The panel is obtained through arranging the visual concepts parsed from the input text by the aid of large language models, and then injected into the denoising network as a detailed control signal to complement the text condition. To facilitate text-to-panel learning, we come up with a carefully designed semantic formatting protocol, accompanied by a fully-automatic data preparation pipeline. Thanks to such a design, our approach, which we call* `Ranni`*, manages to enhance a pre-trained T2I generator regarding its textual controllability. More importantly, the introduction of the generative middleware brings a more convenient form of interaction (*i.e.*, directly adjusting the elements in the panel*

or using language instructions) and further allows users to finely customize their generation, based on which we develop a practical system and showcase its potential in continuous generation and chatting-based editing.

1. Introduction

Language is the most straightforward way for us to convey perspectives and creativity. When we aim to bring a scene from imagination into reality, the first choice is through language description. This forms the philosophical basis of text-to-image (T2I) synthesis. With recent advancements in diffusion models, T2I synthesis demonstrates promising results in terms of high fidelity and diversity [7, 12, 24, 31, 32, 37, 38]. However, the expressive power of language is also limited, compared with structured, pixel-based image in more diverse distribution. This hinders the T2I synthesis to faithfully translate a textual description into a precisely corresponding image. Therefore, current models encounter issues when generating for complex prompts [14], such as determining the quantity of objects, attribute binding, spatial relationship, and multi-subject descriptions.

For professional painters and designers, they express an imagined scene into a tangible form with a broader range of tools beyond just language, *e.g.,* cascading style sheets (CSS) and designing softwares. These tools allow for accurate and enriched expression of visual objects, from the perspectives of spatial positions, sizes, relationships, styles, *etc*. By getting closer to the image modality, they achieve more accurate expression and easier manipulation.

In this paper, our goal is to introduce a new image generation approach, which offers the convenience of text-to-image methods, while also providing accurate expression and enriched manipulation capabilities similar to professional tools. To this end, we present **`Ranni`**, an improved T2I generation framework which translates natural language into a middleware with the help of large language models (LLMs). The middleware, which we call **semantic panel**, acts as a bridge between text and images. It provides accurate understanding of text descriptions and enables intuitive image editing. The semantic panel comprises all the visual concepts that appear in the image. Each concept represents a structured expression of an object. We describe it using various attributes, such as its bounding box, colors, keypoints, and the corresponding textual description.

By introducing the semantic panel, we relax the text-to-image generation with two sub-tasks: *text-to-panel* and *panel-to-image*. During the text-to-panel, text descriptions are parsed into visual concepts by LLMs, which are gathered and arranged inside the semantic panel. The panel-to-image process then encodes the panel as a control signal, guiding the diffusion models to capture the details of each concept. To support efficient training on the above tasks, we present an automatic data preparation pipeline. It extends the text-image pairs of existing datasets by extracting visual concepts using a collection of recognition models.

Based on the semantic panel, `Ranni` also offers a more intuitive way to further edit the generated image. Existing diffusion-based methods [1–3, 25] implicitly understand the editing intention through modified prompts or text instructions. In contrast, we explicitly map the editing intention to an update of the semantic panel. With the rich attributes of visual concepts, we are able to incorporate most editing instructions by composing six unit operations: *addition, removal, resizing, re-position, replacement,* and *attribute revision*. The update of the semantic panel can be done manually through a user interface or automatically by LLMs. In practical, we study the adaption of advanced LLMs on this task. The results demonstrate the potential of a fully-automatic **chatting-based editing** approach, which allows for continuous generation all with text instructions.

2. Related Work

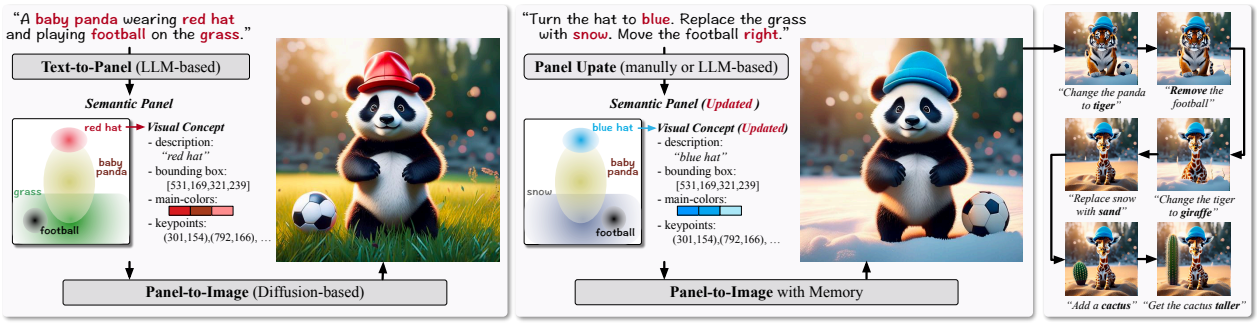
Text-to-Image Generation Models. Diffusion models [7, 12, 24, 34, 37] have become more popular recently compared to GANs [44, 47, 51] and auto-regressive models [8, 9, 32, 33, 45, 46] due to their ability to produce high-quality and diverse outputs. Recent advancements such as Stable Diffusion [38], UnCLIP [31], and Imagen [35] have demonstrated significant improvements in generating text-to-image with an impressive level of photo-realism. `Ranni` builds upon diffusion model, taming it for better instruction following while maintaining the generation quality.

Controllable Generation beyond Texts. Recent works also extend the controllability of diffusion models by including extra conditions such as inpainting masks [15, 43], sketches [40], keypoints [18], depth maps [38], segmentation maps [6, 41], layouts [34], *etc*. Accordingly, models are modified by incorporating additional encoders through either fine-tuning (*e.g*., ControlNet [49], T2I-Adapter [23]) or training from scratch (*e.g*., Composer [15]).

LLM-assisted Image Generation. The large language models (LLMs) have revolutionized various NLP tasks with exceptional generalization abilities, which are also leveraged to assist in text-to-image generation. LLM-grounded Diffusion [19] and VPGen [5] utilize LLMs to infer object locations from text prompts using carefully designed system prompts. LayoutGPT [11] improves upon this framework by providing the LLM with retrieved exemplars. `Ranni` further incorporates a comprehensive semantic panel with multiple attributes. It fully leverages the planning ability of LLMs for accurately following painting and editing tasks.

3. Methodology

We begin by presenting the framework of `Ranni`, which utilizes a semantic panel for accurate text-to-image gen-



(a) Following Painting Instruction **(b) Following Editing Instruction** **(c) Continuous Generation**

Figure 2. The **framework** of Ranni for following painting and editing instructions in a *sequential workflow* based on the semantic panel. **(a)** The painting task is divided into an LLM-assisted text-to-panel, and a diffusion-based panel-to-image generation. **(b)** The editing task is conducted via the update of previous semantic panel. **(c)** The image can be further refined with multi-round compounded editing.

eration. Next, we expand upon the framework to enable interactive editing and continuous generation. The entire framework is depicted in Fig. 2. Lastly, we introduce an automatic data preparation pipeline and the created dataset, which enables the efficient training of Ranni.

3.1. Bridging Text and Image with Semantic Panel

We define the semantic panel as a workspace for manipulating all visual concepts in an image. Each visual concept represents an object, and includes its visually accessible attributes (*e.g.,* position and colors). The semantic panel acts as a middleware between text and image, presenting a structured modeling for text, and a compressed modeling for image. By incorporating the panel, we alleviate the pressure of directly mapping text to image. We include the following attributes for each concept: 1) *text description* for semantic information, 2) *bounding box* for position and size, 3) *main colors* for style, and 4) *keypoints* for shape. The text-to-image generation is then naturally divided two sub-tasks: *text-to-panel* and *panel-to-image*.

Text-to-Panel requires the ability to understand prompts and have a rich knowledge of visual content. We adapt the LLM for this task due to its strong performance as a prompt reader and a task-planner. We design system prompts to request the LLM for imagining the visual concepts corresponding to the input text. When generating multiple attributes of concepts, inspired by chain-of-thought [42], we conduct it in a sequential manner. The whole set of objects is firstly generated with the text descriptions. Detailed attributes, *e.g.,* bounding boxes, are then generated and arranged towards each object. The design of chat templates and examples of full conversations are available in the Supplement Material. Thanks to the zero-shot ability of LLMs, they can generate detailed semantic panels with correct output format. Furthermore, we enhance the performance of LLM by fine-tuning it to better comprehend visual concepts, especially for more detailed attributes like colors. This is achieved by utilizing a large dataset consisting of image-text-panel triples. See Sec. 3.3 for more details.

Panel-to-Image is a task focused on conditional image generation. We implement it using the latent diffusion model [34] as the backbone. To begin, all visual concepts within the semantic panel are encoded into a condition map that has the same shape as the image latent. The encodings of different attributes are as follows:

- *text description*: CLIP text embedding.
- *bounding box*: A binary mask with 1s inside the box.
- *colors*: Indexed learnable embeddings.
- *keypoints*: A binary heatmap with 1s on the keypoints.

These conditions are aggregated using learnable convolution layers. Finally, the condition maps of all objects are averaged to form the control signal.

To control the diffusion model, we add the condition map to the input of its denoising network. The model is then fine-tuned on the dataset described in Sec. 3.3. During inference, we further enhance control by manipulating cross-attention layers of the denoising network. Specifically, for each visual concept, we restrict the attention map of image patches inside its bounding box, giving priority to the words of its text description.

3.2. Interactive Editing with Panel Manipulation

The image generation process of Ranni allows users to access the semantic panel for further image editing. Unlike complex and non-intuitive prompt engineering, editing images using Ranni is more natural and straightforward. Each editing operation corresponds to the update of visual concepts within the semantic panel. Considering the structure of semantic panel, we define the following six **unit operations**: 1) *adding* new objects, 2) *removing* existing ones, 3) *replacing* one with something, 4) *resizing* objects, 5) *moving* objects, and 6) *re-editing* the attributes of objects. Users can perform these operations manually or rely on the assistance of an LLM. For example, "*moving the ball to the left*" can be achieved through a graphical user interface using drag-and-drop, or through an instruction-based chatting procedure with the help of an LLM. We

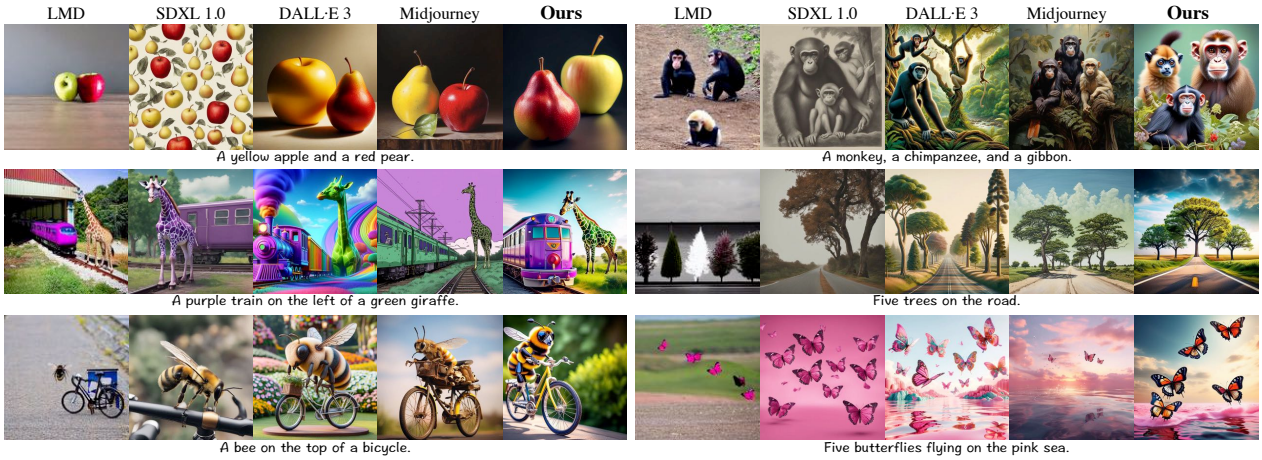


Figure 3. **Comparison on text-to-image generation** between `Ranni` and representative methods.

could also continuously update the semantic panel to refine the image progressively, resulting in more accurate and personalized outputs.

After updating the semantic panel, new visual concepts are utilized to generate the edited image latent. To avoid unnecessary alterations to the original image, we confine the edits to editable regions using a binary mask $\mathbf{M}_e$. Based on the difference between previous and new semantic panel, it is easy to determine the editable areas, *i.e.* bounding boxes of adjusted visual concepts. Assuming the latent representation of the original and current denoising process as $\mathbf{x}_t^{old}$ and $\mathbf{x}_t^{new}$, then the updated representation becomes $\hat{\mathbf{x}}_t^{new} = \mathbf{M}_e \mathbf{x}_t^{new} + (1 - \mathbf{M}_e)\mathbf{x}_t^{old}$.

3.3. Semantic Panel Dataset

To support efficient training of `Ranni`, we build up a fully-automatic pipeline for preparing datasets, consisting of *attribute extraction* and *dataset augmentation*.

Attribute Extraction. We first collect a large set of 50M image-text pairs from multiple resources, *e.g.* LAION [36] and WebVision [17]. For each image-text pair, attributes of all the visual concepts are extracted in the following order:
(i) Description and Box: Grounding DINO [21] is used to extract a list of objects with text descriptions and bounding boxes. We then filter out meaningless descriptions, and remove highly-overlapped boxes with the same description.
(ii) Colors: For each bounding box, we first use SAM [16] to get its segmentation mask. Each pixel inside the mask is mapped into the index of its closest color in a 156-colored palette. We count the index frequency, and pick the top-6 colors with proportions larger than 5%.
(iii) Keypoints: The keypoints are sampled within the SAM [16] mask using the FPS algorithm [30]. Eight points are sampled, with an early stopping when the farthest distance of FPS reaches a small threshold of 0.1.

Dataset Augmentation. We empirically find it efficient to augment the dataset with following strategies:
(i) Synthesised Captions: The original caption of an image might ignore some objects, resulting in incomplete semantic panels. To address this issue, we utilize LLaVA [50] to find out images with multiple objects (in total of 2M) and generate more detailed captions for them.

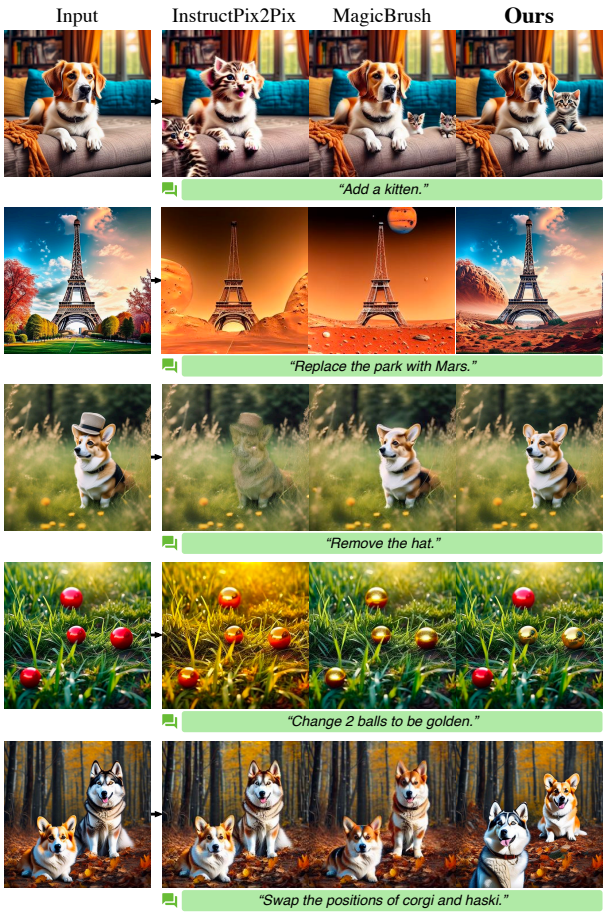


Figure 4. **Comparison on instruction editing** between `Ranni` and representative methods, using unit operation prompts.

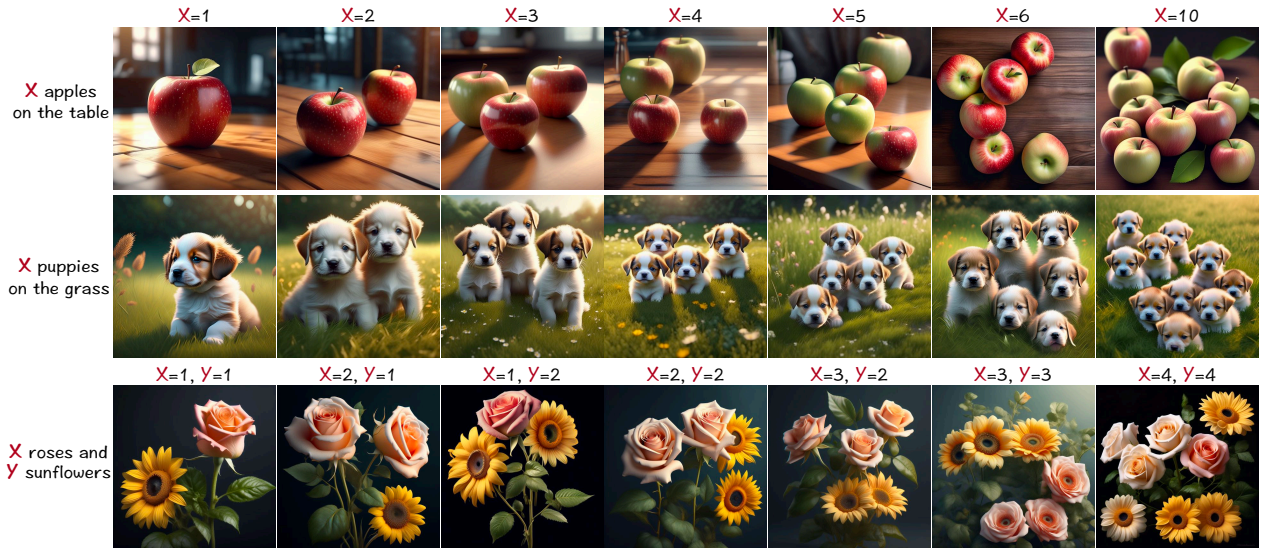


Figure 5. Samples generated by Ranni on **quantity-awareness** prompts.

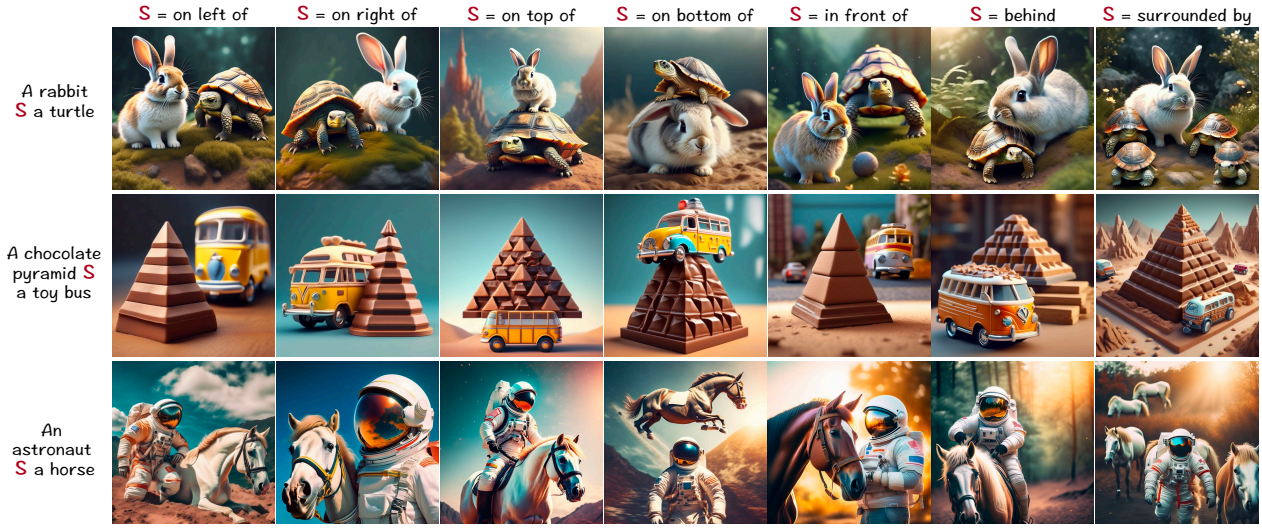


Figure 6. Samples generated by Ranni on **spatial relationship** prompts.

(ii) Mixing Pseudo Data: To enhance the ability of spatial arrangement, we create pseudo samples using manual rules. We generate random prompts from a pool of objects with varying orientations, colors, and numbers. Next, we synthesize the semantic panel by arranging them randomly according to specified rules.

4. Experiments

4.1. Experimental Setup

For the **text-to-panel** task, we select the open-sourced Llama-2 [39] 13B version as our LLM. To enable attribute generation for each parsed object, we fine-tune the LLM with LoRA [13] for 10K steps with a batch size of 64. The final optimized module for each attribute generation task contains 6.25M parameters, making it easy to switch between tasks. The datasets are sampled with a mixture of 50% probability from subsets with raw captions, 45% from synthesized captions, and 5% from pseudo data.

For the **panel-to-image** task, we fine-tune a pre-trained latent diffusion model with 3B parameters on our constructed dataset with visual concepts. The fine-tune process contains 40K steps with a batch size of 128. Training samples are evenly distributed between raw and synthesised captions. To prioritize attribute conditions over text conditions in the model, we apply a drop rate of 0.7 to the text conditioning.

4.2. Evaluation on Text-to-Image Alignment

Qualitative Evaluation. We expect Ranni to generate images that align better with text. In this section, we examine its alignment ability with various types of prompts,

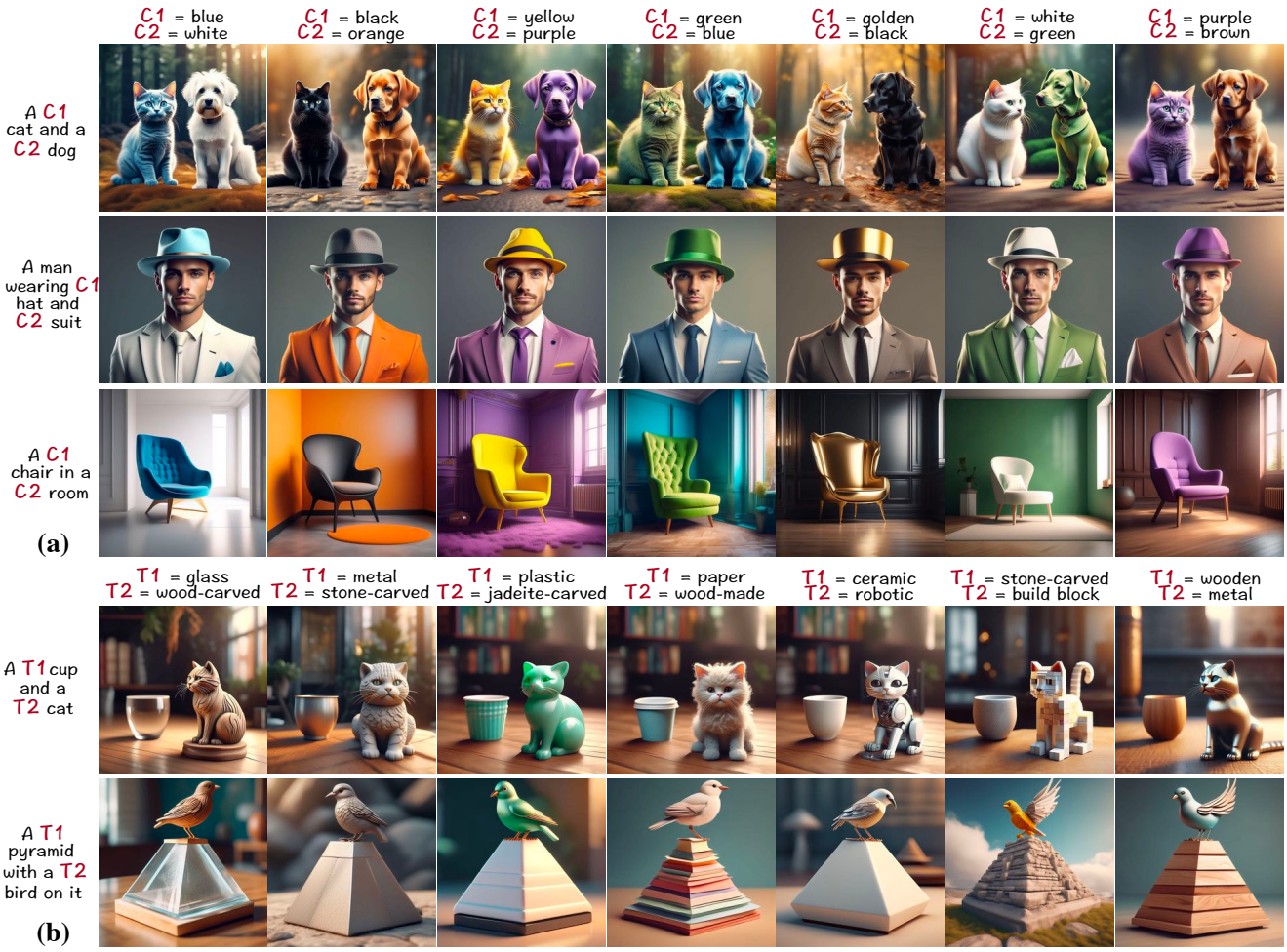


Figure 7. Samples generate by `Ranni` on **attribute binding** prompts, including the (a) color binding and (b) texture binding. For clear comparison, the random seed is fixed to preserve the spatial arrangement in one row.

Figure 8. Samples generated by `Ranni` on **multi-object** prompts.

which are known to be challenging for existing methods:

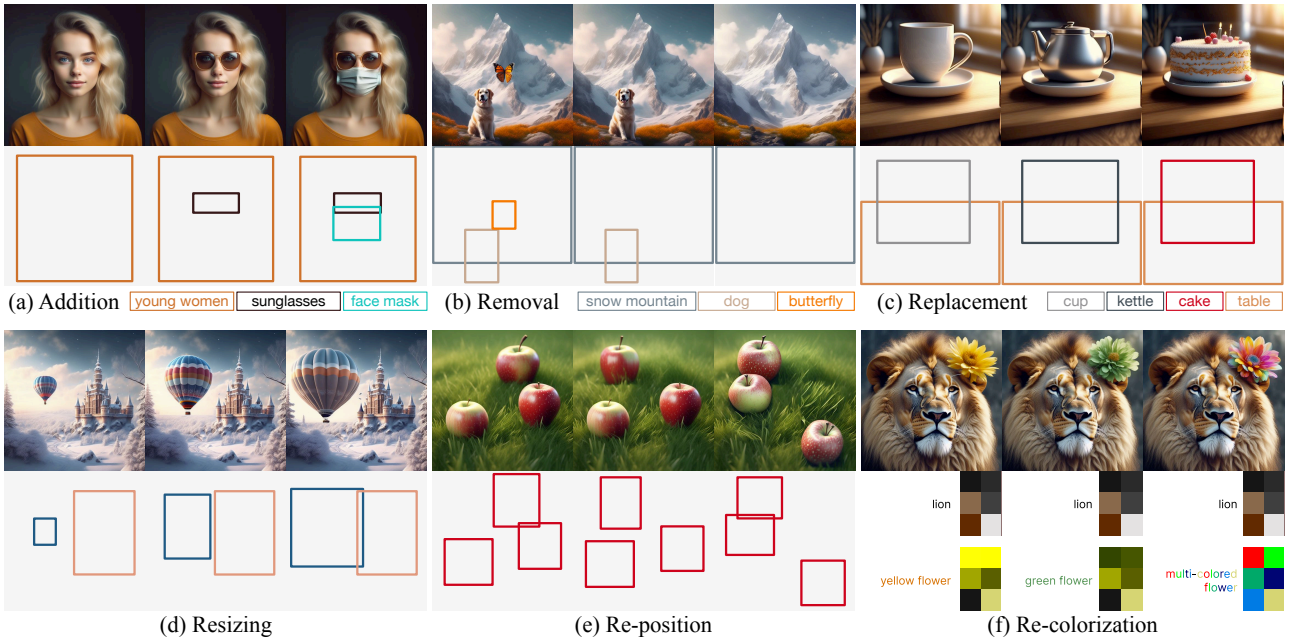
Quantity: Existing model struggles to generate objects in the exact requested number. Fig. 5 shows that `Ranni` is more sensitive to the varying numbers of objects.

Spatial Relationship: We examine the spatial awareness of `Ranni` with 7 types of relationships in Fig. 6. Results show its ability to properly arrange object positions.

Attribute Binding: We show cases in Fig. 7 on objects with varying attributes. `Ranni` explicitly distinguishes different objects, thus allowing for precise assignment of attributes to each object without any cross-influence.

Multiple Objects: When generating multiple objects with a similar appearance, existing models might confuse them together. Fig. 8 shows that `Ranni` successfully generates groups containing similar people, animals or plants.

Quantitative Evaluation. We further evaluate the alignment with quantitative metrics. For attribute binding and spatial relationship, we use the validation prompts and metrics from T2I-CompBench [14], with 300 prompts for each subset. For quantity, we generate 300 prompts containing 1 to 5 objects in same type. We set the score as the proportion of results that generate correct number of objects. For multiple objects, we start by collecting 30 groups, each



(d) Resizing (e) Re-position (f) Re-colorization

Figure 9. The editing results and corresponding panel update for each **unit operation**.

containing 4 similar objects such as tiger, lion, cat, and leopard. Next, for each group, we generate 10 prompts that include 2 to 4 different objects. The metric used is the BLIP-VQA and UniDet score for attribute binding and spatial relationship, respectively [14]. For quantity and multi-object, we extend BLIP-VQA with more than two questions, based on the exact number of objects.

Tab. 1 shows the evaluation results. `Ranni` outperforms existing methods, including end-to-end models and inference-optimized strategies. In particular, it shows great improvement on the spatial relationship and quantity-awareness tasks. We also compare with our pre-trained base model. The improvement suggests that `Ranni` could enhance the prompt following of an existing model with semantic panel control.

Visualized Comparison. To compare with existing models, we visualize the results on different types of prompt in Fig. 3. We compare `Ranni` with LLM-grounded diffusion (LMD) [19], Stable Diffusion XL 1.0 [29], DALL·E 3 [27], and Midjourney [22]. `Ranni` achieves competitive performance in prompt following, while maintaining the fidelity of its generation. It is noteworthy that `Ranni` demonstrates improved alignment in terms of quantity-awareness and spatial relationship, which is consistent with the quantitative results in Tab. 1.

4.3. Evaluation on Interactive Generation

Based on the generated image with its semantic panel, `Ranni` can make further edit to the image at a high semantic level. We first evaluate `Ranni`'s performance on unit editing operations. Then we expand its capabilities to include multi-round editing with compounded operations. Lastly, we enhance `Ranni` by incorporating the intelligence of LLMs to enable chatting-based editing.

Table 1. **Quantitative results for alignment assessment** on various benchmarking subsets. The best and second results for each column are **bold** and underlined, respectively.

Method	Attribute Binding			Spatial	Quantity	Multi-obj.
	Color	Texture	Shape			
SD v1.5 [38]	0.3730	0.4219	0.3646	0.1312	0.1801	0.4255
SD v2.1 [38]	0.5694	0.4982	0.4495	0.1738	0.2337	0.5562
Composable [20]	0.4063	0.3645	0.3299	0.0800	-	-
Structured [10]	0.4990	0.4900	0.4218	0.1386	-	-
Attn-Exct [4]	0.6400	0.5963	0.4517	0.1455	-	-
GORS [14]	0.6603	0.6287	0.4785	0.1815	-	-
SDXL (b2r) [29]	0.6050	0.5446	0.4780	0.2086	0.1992	0.5905
SDXL (bpr) [29]	0.6132	0.5331	0.4896	0.2097	0.1839	0.6221
Base model	0.5446	0.5970	0.4732	0.1833	0.2337	0.5579
`Ranni` (Ours)	**0.6893**	**0.6325**	**0.4934**	**0.3167**	**0.2720**	**0.6400**

Unit Operation defined in Sec. 3.2 is the basis for all editing operations. Most editing intentions can be considered as one or a combination of the unit operations. Fig. 9 shows the correspondence between each unit operation and the update of semantic panel. We compare the editing ability w.r.t. unit operations with Instruct-Pix2Pix [3] and MagicBrush [48] in Fig. 4. We can see that `Ranni` could better preserve the non-editing area, and achieve more flexible operations, *e.g.* swapping positions.

Compounded Operations. Based on the unit operations, we further apply `Ranni` for continuous editing with compounded operations. In Fig. 10, we present examples of progressively creating images with complex scenes. During this interactive creation process, users can refine the image

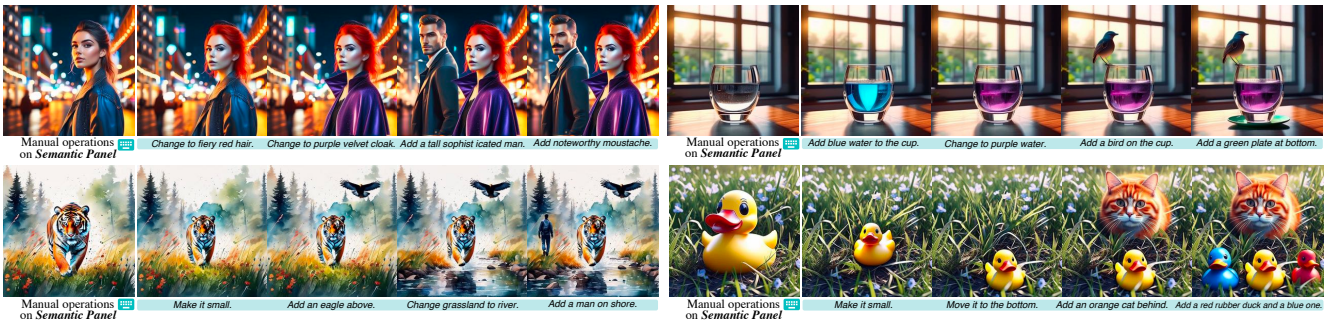


Figure 10. Results of **continuous generation** with multi-round editing chains, consisting of unit operations.

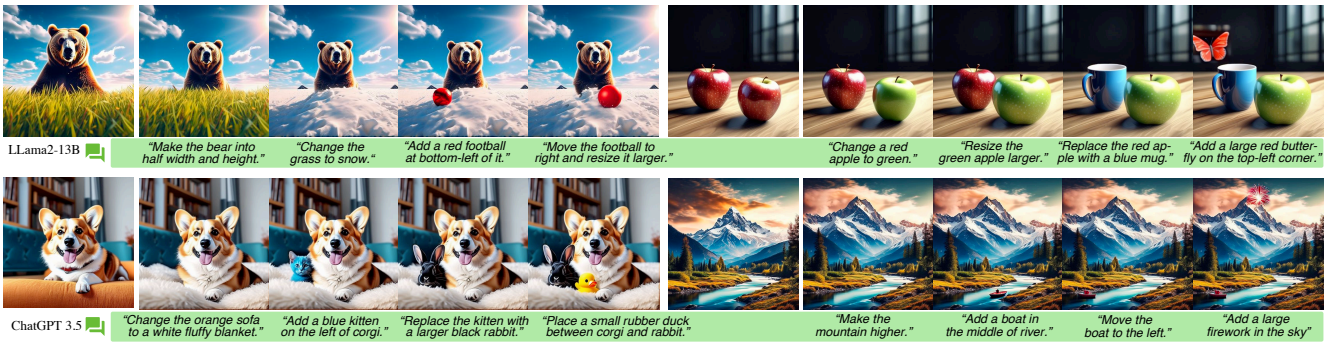


Figure 11. Results of **chatting-based generation** in natural instructions with different LLMs . Refer to Fig. 1 for the results of ChatGPT-4.

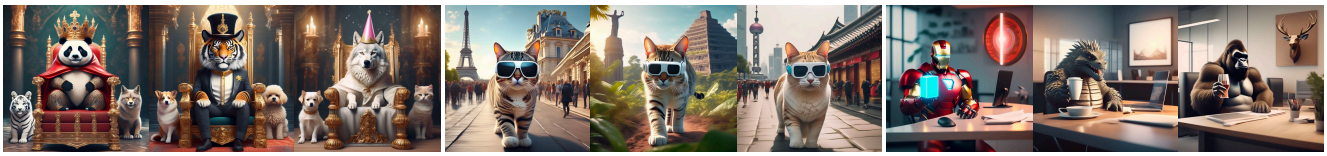


Figure 12. Samples generated by `Ranni` with **similar layouts**.

step-by-step by replacing unsatisfying objects, adding more details, and experimenting with various attributes. The interactive nature also enables additional applications, such as generating images with similar layouts, as shown in Fig. 12. To achieve this, we first generate an image as the base and then sequentially replace objects in it.

Chatting-based Editing. We also use LLMs to automatically map editing instructions into updates of the semantic panel. To accomplish this, we introduce new system prompts that are specifically designed for this task. These system prompts request the LLM to understand the current semantic panel and the editing instruction, and then generate the updated panel. Please refer to the Supplementary Material for more details. Fig. 11 and Fig. 1 (c) present cases using Llama2-13B [39], ChatGPT-3.5 [26], ChatGPT-4 [28], respectively. During the evaluation, we observed that LLM has the ability to understand more natural instructions. For instance, the instruction "*the mushroom is eaten*" indicates the need to remove it, while "*the mushroom grows higher*" implies increasing its height while keeping its bottom position intact. The results demonstrate the potential of `Ranni` as a unified image creation system that supports sequential instructions with chatting.

5. Conclusion

In this paper, we present `Ranni`, a new approach that tames existing diffusion models to better follow the painting and editing instructions. The semantic panel in `Ranni` is introduced as a generative middleware between text and image. It helps relieve the pressure of directly mapping complex prompt to image. The panel is firstly constructed using visual concepts parsed by LLM from the given prompt. It then serves as control signal to complement the generation of diffusion models. `Ranni` follows painting instruction without ignoring detailed description of each concept in prompt. Furthermore, by adjusting the semantic panel with manual or LLM-based operations, `Ranni` enables interactive editing of previous generated images. We demonstrates that with the fully automatic control of LLM, `Ranni` shows potential as a flexible chat-based image creation system, where any existing diffusion model can be incorporated as the generator for interactive generation.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18187–18197, 2022. 2

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Eur. Conf. Comput. Vis.*, volume 13675, pages 707–723, 2022.

[3] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18392–18402, 2023. 2, 7

[4] 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 (TOG)*, 42(4):1–10, 2023. 7

[5] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *ArXiv*, abs/2305.15328, 2023. 2

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *ArXiv*, abs/2210.11427, 2022. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Adv. Neural Inform. Process. Syst.*, 34:8780–8794, 2021. 2

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering text-to-image generation via Transformers. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[9] Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for high-resolution image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12868–12878, 2020. 2

[10] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *Int. Conf. Learn. Represent.*, 2023. 7

[11] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *ArXiv*, abs/2305.15393, 2023. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.*, 33:6840–6851, 2020. 2

[13] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent.*, 2021. 5

[14] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 2, 6, 7

[15] Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. Composer: Creative and controllable image synthesis with composable conditions. In *Int. Conf. Mach. Learn.*, 2023. 2

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *ArXiv*, abs/2304.02643, 2023. 4

[17] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. *ArXiv*, abs/1708.02862, 2017. 4

[18] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22511–22521, 2023. 2

[19] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *ArXiv*, abs/2305.13655, 2023. 2, 7

[20] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Eur. Conf. Comput. Vis.*, volume 13677, pages 423–439, 2022. 7

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023. 4

[22] midjourney. Midjourney, 2023. 7

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *ArXiv*, abs/2302.08453, 2023. 2

[24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Int. Conf. Mach. Learn.*, pages 8162–8171. PMLR, 2021. 2

[25] Alexander Quinn 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. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *Int. Conf. Mach. Learn.*, volume 162, pages 16784–16804, 2022. 2

[26] OpenAI. ChatGPT, 2023. 8

[27] OpenAI. DALL·E 3, 2023. 7

[28] OpenAI. GPT-4 technical report. *ArXiv*, abs/2303.08774, 2023. 8

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint*

arXiv:2307.01952, 2023. 7

[30] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Adv. Neural Inform. Process. Syst.*, pages 5099–5108, 2017. 4

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 2

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Int. Conf. Mach. Learn.*, pages 8821–8831. PMLR, 2021. 2

[33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 2, 3

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

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Adv. Neural Inform. Process. Syst.*, volume 35, pages 25278–25294, 2022. 4

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 2

[38] stability.ai. Stable Diffusion 2.0 Release, 2022. 2, 7

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. 5, 8

[40] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[41] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *ArXiv*, abs/2205.12952, 2022. 2

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Adv. Neural Inform. Process. Syst.*, 2022. 3

[43] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22428–22437, 2023. 2

[44] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1316–1324, 2018. 2

[45] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. *ArXiv*, abs/2110.04627, 2021. 2

[46] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv*, abs/2206.10789, 2022. 2

[47] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 833–842, 2021. 2

[48] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *ArXiv*, abs/2306.10012, 2023. 7

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, 2023. 2

[50] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *ArXiv*, abs/2306.17107, 2023. 4

[51] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5802–5810, 2019. 2