Omni-Dish: Photorealistic and Faithful Image Generation and Editing for Arbitrary Chinese Dishes

Huijie Liu[1,2], Bingcan Wang[1], Jie Hu[1], Xiaoming Wei[1], Guoliang Kang[2]
[1] Meituan, China
[2] Beihang University, China

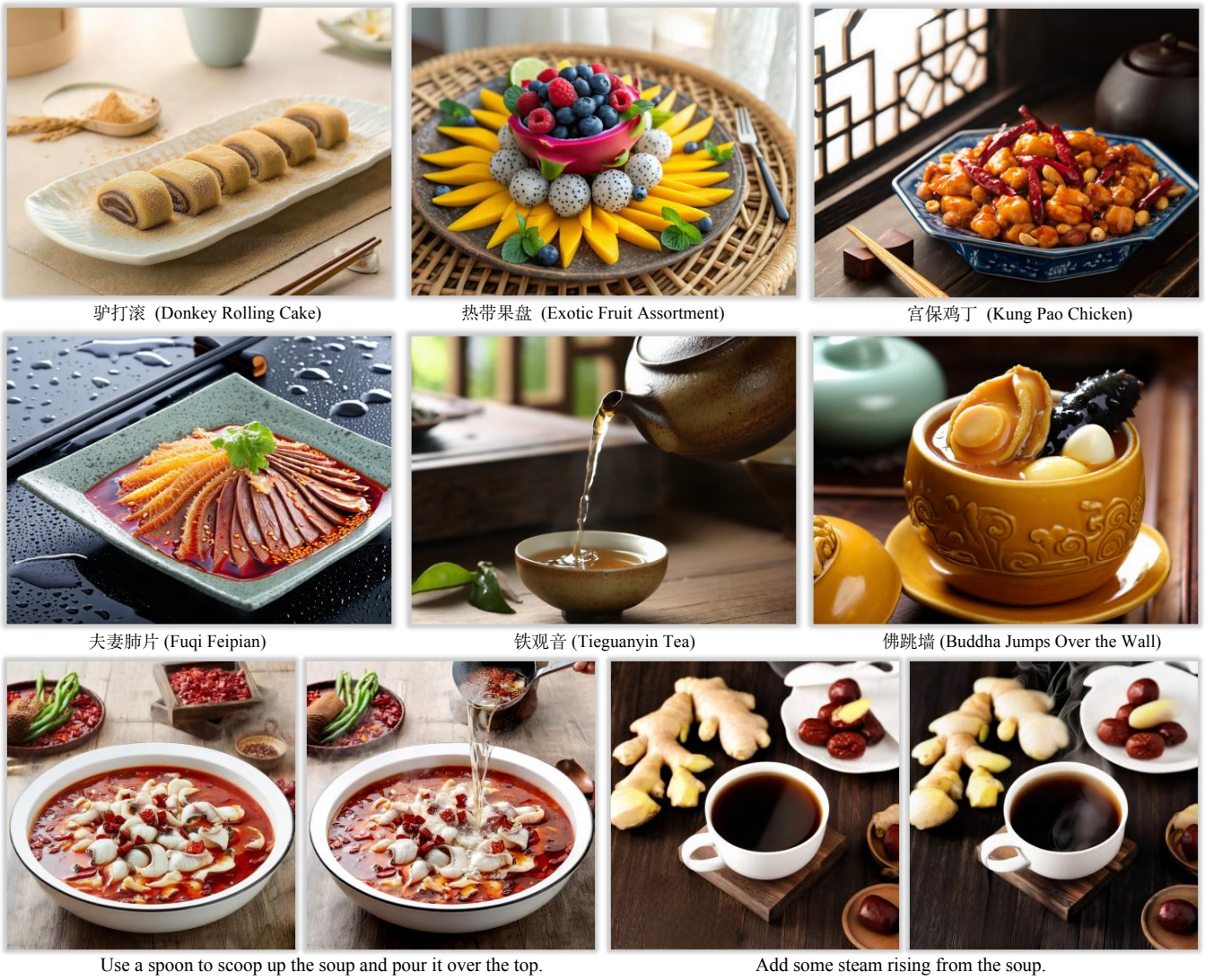

Figure 1: Samples demonstrate the superiority of our method in generating and editing Chinese culinary dishes. Rows 1-2 show the faithful generation results from Omni-Dish. Row 3 shows the dish editing capabilities. Detailed descriptions and authentic photos of all referenced Chinese dishes can be found in the Appendix.

Abstract

Dish images play a crucial role in the digital era, with the demand for culturally distinctive dish images continuously increasing due to the digitization of the food industry and e-commerce. In general cases, existing text-to-image generation models excel in producing high-quality images; however, they struggle to capture diverse characteristics and faithful details of specific domains, particularly Chinese dishes. To address this limitation, we propose **Omni-Dish**, the first text-to-image generation model specifically tailored for Chinese dishes. We develop a comprehensive dish curation pipeline,

building the largest dish dataset to date. Additionally, we introduce a recaption strategy and employ a coarse-to-fine training scheme to help the model better learn fine-grained culinary nuances. During inference, we enhance the user's textual input using a pre-constructed high-quality caption library and a large language model, enabling more photorealistic and faithful image generation. Furthermore, to extend our model's capability for dish editing tasks, we propose Concept-Enhanced P2P. Based on this approach, we build a dish editing dataset and train a specialized editing model. Extensive experiments demonstrate the superiority of our methods.

CCS Concepts

• **Computing methodologies** → **Computer vision**; *Appearance and texture representations*; Image representations; • **Information systems** → **Multimedia content creation**.

Keywords

Generative Models, Diffusion Models, Image Generation, Image Editing, Chinese Dishes

1 Introduction

Dishes are closely connected to daily human experiences, and dish images are becoming an important medium in the digital era. High-quality dish images are particularly indispensable in daily lives. Consequently, generating photorealistic and faithful dish images through text input has a significant impact on multiple industries.

As shown in Fig. 2, existing text-to-image models [21, 27, 34, 40, 42] demonstrate some limitations: (1) arbitrariness: unable to recognize dish names with Chinese cultural significance; for example, some models can not recognize that "Buddha Jumps Over the Wall" is a kind of Chinese dish; (2) photorealism and faithfulness: unable to capture nuanced details; for example, they are unable to clearly render the details and textures of "Donkey Roll Cake".

To overcome this limitation, we propose **Omni-Dish**, a text-to-dish generation model, designed for arbitrary photorealistic and faithful dish generation. To achieve **arbitrary dish generation**, including various niche dishes, we build an unparalleled large-scale dataset. We collect 100 million dish name-image pairs from China's largest catering website, subsequently implementing a meticulously designed data curation pipeline. The pipeline employs Large Language Models (LLMs) [1, 12, 51, 52], Vision Language Large Models (VLLMs) [3, 9, 45, 48], an aesthetic scoring model, and other specialized models to perform data filtering, correction, and tagging.

However, we find that scaling up the dataset is insufficient for **photorealistic and faithful generation**. Thus, we propose a recaption and rewriting strategy. For recaption, we find that it is not suitable to directly use VLLMs to caption dish images as in previous methods [13, 37], because they struggle to accurately identify the subtle elements in dish images. Therefore, we first use LLMs [5] to describe the dishes. These descriptions are then paired with images and fed into VLLMs for recaption, which acquires more faithful descriptions. With these recaptions, we implement a coarse-to-fine training pipeline. In the initial phase, the model is trained with dish images and dish names (without VLLMs' recaptions) to learn foundational dish concepts. In the subsequent phase, dish images accompanied by high-quality recaptions are leveraged to capture fine-grained representations. During inference, we enhance user input with a pre-constructed high-quality caption library, and further rewrite the captions with LLMs to improve the quality of generated images. Our recaption and rewriting strategy not only enables the model to generate faithful dish images, but also improves its ability to follow instructions, as shown in Fig. 3.

Figure 2: Existing methods face challenges in generating photorealistic and faithful images of arbitrary Chinese dishes. Row 1 shows reference images that are real photographs.

Building upon our dish generation model, we further develop a **dish editing** pipeline that demonstrates the generation model's capacity to enable diverse downstream tasks. Existing image editing methods [4, 22, 23, 55, 56] struggle to achieve dish editing, because (1) editing ingredients in dishes is often very subtle, making it difficult to achieve these fine-grained changes; (2) some editing types for dishes are highly customized, as shown in row 3 of Fig. 1.

Following [4], we construct a dataset using Prompt-to-Prompt (P2P) [16]. We find that there exists a trade-off in the attention replacement steps of P2P, as shown in Fig.6. Extensive step replacement leads to compromised consistency in generated data pairs, while limited step replacement results in compromised consistency. Thus, we propose Concept-Enhanced P2P, which leverages the generation model to generate concept-specific images (not image pairs) and fine-tunes the generation model using these generated images to amplify its generation capability of the specific concept. Then, we use P2P to generate image pairs with the fine-tuned model. In addition, we construct data pairs for the addition and removal of ingredients in dishes by inpainting. After final human filtering, we acquire a dataset with high consistency and observable editing effects and use it to train a dish editing model.

In the end, we conducted extensive experiments to evaluate our approach. For the first time, we introduced the use of the Dish T2I similarity [49] to evaluate the model. Furthermore, we constructed a high-aesthetic dish dataset as test data for FID [17]. We also conducted comprehensive human evaluations across multiple dimensions. Specifically, we evaluated fidelity, texture, composition, scene, lighting, and subject for dish generation, as well as the effectiveness, consistency, and aesthetics for dish editing. Extensive experiments demonstrate the effectiveness of our method.

In summary, our contributions can be summarized as follows:

"扬州炒饭，金黄的米粒在青瓷莲花碗中堆成半球形，虾仁、火腿丁与青豆如星点镶嵌其中。配套汤匙以45度角斜靠碗沿，背景的蓝白扎染桌布与青瓷形成对比，窗边自然光让每粒米饭都泛着光泽。"
"Yangzhou fried rice: golden rice grains form a hemispherical shape in a celadon lotus bowl. Shrimp, diced ham, and green peas are embedded like star-like specks. A matching spoon leans against the bowl's rim at a 45-degree angle. The blue-and-white tie-dyed tablecloth in the background contrasts with the celadon. Natural light from the window makes each rice grain glisten."

"牛肉面，盛放在青花瓷碗中被筷子夹起，琥珀色清汤里浮着半透明萝卜片与红色辣油，香菜碎在汤面聚成翠色岛屿。几片牛腱肉沿着碗边扇形铺开，背景的开放式厨房中，木质案板上的面粉在侧光下如初雪闪烁。"
"Beef noodles: severd in a blue-and-white porcelain bowl held by chopsticks, clear amber broth floats translucent radish slices and red chili oil. Chopped cilantro forms green islands on the soup's surface. Beef shank slices fan out along the bowl's rim. In the background open kitchen, flour on a wooden board glistens like snow under side lighting."

"猪肉炖粉条，盛放在深褐陶钵内，五花肉块与水晶粉条交错层叠，葱花碎在油润汤面聚成星点翠色，蒸汽腾腾升起。背景是农家土灶台柴火余烬泛着暗红。"
"Pork stew with vermicelli: in a deep brown earthen pot, pork belly pieces and crystal noodles interlace in layers. Chopped scallions gather into star-like green specks on the oily soup surface, with steam billowing upward. The background shows dark red glowing embers in a rural earthen stove's firewood ashes."

Figure 3: Nuanced descriptions not only help Omni-Dish generate faithful dish images, but also endowing it with fine-grained instruction-following capabilities.

- We are the first to propose an image generation model specifically for dishes, **Omni-Dish**. We introduce a novel dish data curation pipeline and a recaption and rewriting strategy for training and inference, capable of generating arbitrary photorealistic and faithful dish images.
- We extend Omni-Dish's capability for supporting dish editing. Building upon Omni-Dish, we present Concept-Enhanced P2P for constructing the first open source dish editing dataset, which enables the training of editing models.
- We conducted extensive experiments, including automated metrics and carefully designed human evaluations, which demonstrate the superiority of our approach.

2 Related Work

Image Generation. Text-to-Image generation aims to generate images conditioned on text input. Previous researchers focused on (GANs) [2, 15, 24, 25], which have gradually been replaced by more advanced models [10, 37, 42]. The Denoising Diffusion Probabilistic Model [18] proposed diffusion models based on U-Net [41], which inspired some methods such as Stable Diffusion [34, 40]. Recently, some models such as PixArt [7, 8] and FLUX [27] replace U-Net with DiT [32]. However, these models, trained with English prompts, struggle to process Chinese dish names. Even Chinese-capable models [21, 44] such as Seedream 2 [14] and Cogview [46, 57] have difficulty generating faithful Chinese dish images.

Instruction-Based Image Editing. Instruction-based image editing involves altering images based on instructions, with the key challenge being the construction of high-quality datasets for model training. HQ-Edit [23] constructs datasets by having DALL·E 3 [13] create diptychs. However, these diptychs often exhibit poor consistency. AnyEdit [53] uses SAM, inpainting models [29] and PIH [47] to construct a dataset. Some other methods [4, 22, 56] use Prompt-to-Prompt (P2P) [16] to generate image pairs. P2P represents a prevalent paradigm in image editing, although it inherently lacks image manipulation capabilities for existing photographs. Most editing methods [11, 28, 31, 39, 43, 53, 54] incorporate diverse designs for datasets, but often lack specialized data for dish editing.

3 Omni-Dish: Text-to-Dish Generation

In this section, we propose **Omni-Dish**, the first text-to-dish generation model. Text-to-Dish generation refers to generating photorealistic and faithful dish images based on textual dish names. With our main emphasis on Chinese dishes, Omni-Dish can accept Chinese text input. We present our methodology in three aspects: data curation, model training and inference.

3.1 Data Curation Pipeline

To fully cover all Chinese dish concepts, we collect 100 million data entries from China's largest catering website to build the raw dataset, which is composed of dish name-image pairs.

Data Filtering. We first use VLLMs to detect and filter out images containing text, watermarks, and human hands. Subsequently, we use a dish detection model [50] to localize dish items through bounding box annotations, where samples with bounding boxes exceeding image boundaries (indicating incomplete dish presentation) are rigorously excluded.

Data Correction. The data correction pipeline is illustrated in Fig. 5. To tackle the issue of prevalent noisy dish names, we implement LLMs to: (a) filter out irrelevant entries that lack semantic relevance to dishes (Column 1 of Fig. 5); (b) correct overly descriptive but valid names (Columns 2-4 of Fig. 5). Unfortunately, we find that the dish names corrected by LLMs are not always the correct ones we desire (Column 3, Fig. 5). To address this limitation, we propose to implement dish text-to-image similarity (DishSim.) [50] to quantitatively evaluate the semantic alignment between candidate dish names (original vs. corrected) and their corresponding images. The naming variant that demonstrates the superiority of DishSim. is selected as the final correct dish name.

Data Tagging. We annotate the data with multi-dimensional tags encompassing four aspects: aesthetic quality, tableware, background, and camera angle. For example: "served in a white ceramic bowl, placed on a brown wooden tabletop, high aesthetic quality,

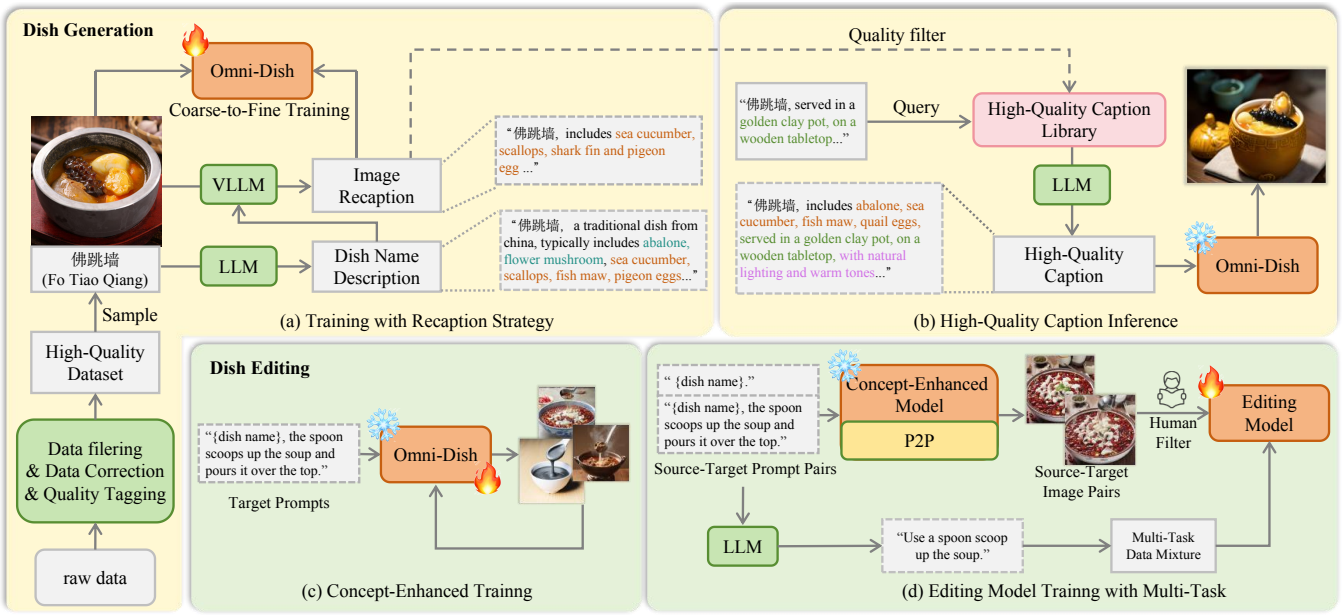

Figure 4: Overview of our method. In the yellow block, (a) with the dish curation and recaption, the coarse-to-fine strategy is applied to train Omni-Dish; (b) high-quality captions are obtained from a pre-constructed library and rewritten by large language models for inference. In the green block, (c) the Concept-Enhanced P2P approach is introduced to build the dish editing dataset; (d) a dish editing model is trained through a multi-task data mixture.

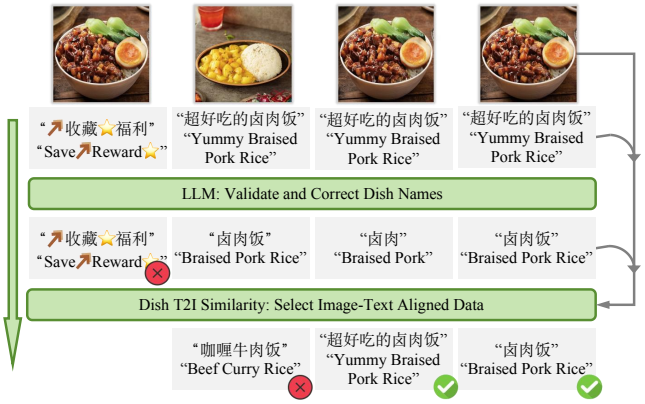

Figure 5: Dish name correction by two steps. For details, refer to Data Correction in Sec. 3.1.

30-degree shooting angle." These tags are used for model training as extra textual input to enhance instruction-following capabilities.

3.2 Faithful-Enhanced Training and Inference

Two-Stage Recaption Strategy. Unlike most common images encountered in daily life, dish images demand greater fidelity to capture faithful details such as ingredients and textures. We observe that models struggle to learn dish-specific nuances and textures when relying solely on dish names as textual input. To overcome this difficulty, a straightforward approach is to employ vision large language models (VLLMs) [3, 9, 45, 48] for image recaption. However, VLLMs often fail to infer granular information about fine-grained material composition or cooking methods directly from visual input. Consequently, we first employ a Large Language Model (LLM) [1, 12, 51, 52] to generate comprehensive generic descriptions of dishes using dish names (rather than dish name-image pairs). Then, VLLMs can generate fine-grained and high-fidelity recaptions for dish images based on the introduction provided by the LLMs as a prior.

Coarse-to-Fine Training. The training objectives can be decoupled into two components: dish concept learning and image quality enhancement. During the concept learning phase, we find that directly training the model with dish names and captions led to convergence difficulties. Therefore, we first train the model using standalone dish names combined with tags generated during data curation. In the subsequent phase, we incorporate more complex recaptions. In the image quality enhancement phase, we employ manually annotated ultra-high-quality data for fine-tuning to improve the aesthetic quality. Furthermore, we use annotated human preference data for Direct Preference Optimization (DPO) [36] to enhance the stability of image generation. In summary, we employ a coarse-to-fine training strategy as illustrated in Tab. 1.

High-Quality Caption Inference. When users interact with the model, they often cannot provide rich fine-grained textual input like recaptions mentioned before, but instead only input a dish name with minimal description. This does not fully utilize the model's capability to generate details learned from recaptions. Therefore, we filter high-quality image data from the training set and generate recaptions using the aforementioned method to pre-construct a high-quality caption library, which contains data entries composed of dish names, high-quality captions, and CLIP embeddings. The library covers nearly all Chinese dishes with multiple captions for each dish. During inference, given a user-input dish name, we

Table 1: Coarse-to-fine training. "Tags" refers to textual description across 4 dimensions (Sec. 3.1). All data in Stage 4 and Stage 5 undergo manual annotation.

Stage	Sample	Resolution
Stage 1	Dish name + Tags	512
Stage 2	Dish name + Tags + Recaption	512
Stage 3	Dish name + Tags + Recaption	1024
Stage 4	Dish name + Tags + Recaption (Ultra-High Quality)	1024
Stage 5	Human Preference Data (for DPO)	1024

first query the library for captions associated with that dish and then calculate the CLIP similarity. After that, we feed both the caption with the highest similarity and the user's input text to the LLMs, instructing them to rewrite the caption to align with the user's description. In this way, we obtain high-quality dish captions that meet user requirements, as illustrated in Fig. 3. In addition, we supplement these captions with high quality tags (e.g., "high aesthetic quality", "high definition"; see Sec. 3.1). Ultimately, these captions will be fed into Omni-Dish as inference inputs.

Model Structure. Omni-Dish adopts the FLUX [27] architecture with 7B parameters, while keeping the VAE of FLUX.1-dev [27] frozen and optimizing the denoising backbone. To support the Chinese language, we use Qwen2.5-7B [52] as a text encoder. Although our implementation primarily focuses on this architecture, the proposed method is broadly applicable to various diffusion model structures such as U-Net [41] and DiT [32].

4 Instruction-Based Dish Editing

Instruction-based dish image editing is more challenging compared to general editing tasks. Existing methods [4, 22, 23, 31, 53–55] struggle with dish image editing. This is because the elements in dish images are more fine-grained, and instructions for dishes are more complex. To further demonstrate the value of Omni-Dish as a dish foundation model, we utilize it to construct a dish editing dataset, and subsequently train an editing model tailored for dishes.

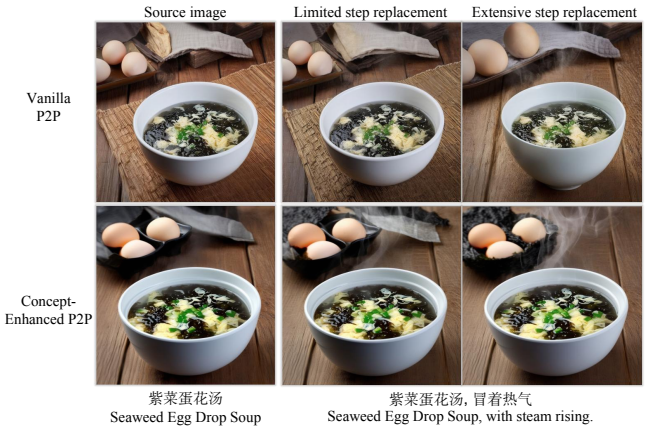

Figure 6: Concept-Enhanced P2P can enhance editing effects while maintaining consistency.

4.1 Concept-Enhanced P2P

Prompt-to-Prompt (P2P) [16] is a training-free method that directly maps text edits to the image generation process by replacing cross-attention maps. Given two paired prompts, it can generate paired images with consistent structure. Existing instruction-based editing methods [4, 55] typically leverage LLMs to construct paired prompts (source and target prompts). Then they use these prompts to generate paired images by applying P2P to T2I foundation models, which heavily relies on the capabilities of foundation models. Omni-Dish can generate high-quality dish images, allowing us to use it for creating high-quality editing data.

However, we observe a trade-off between consistency preservation and editing effectiveness in P2P. P2P achieves its editing capability by replacing attention maps at specific timesteps during the denoising process. In this operation, excessive replacement of timesteps degrades output consistency, while insufficient timestep replacement limits editing flexibility. As shown in Fig. 6, vanilla P2P is almost unable to generate steam when fewer steps are replaced; when more steps are replaced, weak steam is generated, but there are noticeable inconsistencies. It is due to the model's shallow fitting of the "steam" concept. Therefore, we first use Omni-Dish to generate images conditioned on various types of target prompts. Then use them to fine-tune Omni-Dish, thereby enhancing its ability to fit the target concept. To help the model better learn the concept of editing types, we also incorporate some data generated from source prompts. Ultimately, we apply P2P to this concept-enhanced Omni-Dish, which successfully builds dataset for complex editing types.

It is worth noting that we are not solely relying on P2P to build the dataset. We use VLLM to identify elements in the image suitable for removal, such as peppers and coriander, then use Grounded-SAM [26, 30, 38] to find the mask of the element and remove the element by inpainting. We utilize these data bidirectionally to construct both "remove" and "add" type editing data. To ensure the quality of the training data, all data undergo human filtering.

4.2 Dish Editing Training and Inference

During training, we find that it is more critical for the model to learn the execution of editing operations (e.g., remove, add) rather than merely acquiring editing elements. For instance, training the editing model solely on "add steam" dish data (while excluding other editing-operation data) causes catastrophic failure in comprehending the general "add" operation, resulting in severely degraded output quality. To mitigate this, we augment dish data with multi-task data mixture. Specifically, we incorporate editing data from general scenarios into the dish editing dataset to assist the model in understanding editing operations.

To incorporate the source image as an image condition, we concatenate tokens from patchified embeddings of both the noisy latent and the source image as model input. We augment the generation foundation model with three randomly initialized channels to process source image information. Employing a DiT [32] architecture, we train an editing model using paired data generated through the Omni-Dish framework. After training, the model demonstrates the capability to produce well-aligned edited images that maintain consistency with both textual instructions and the source image.

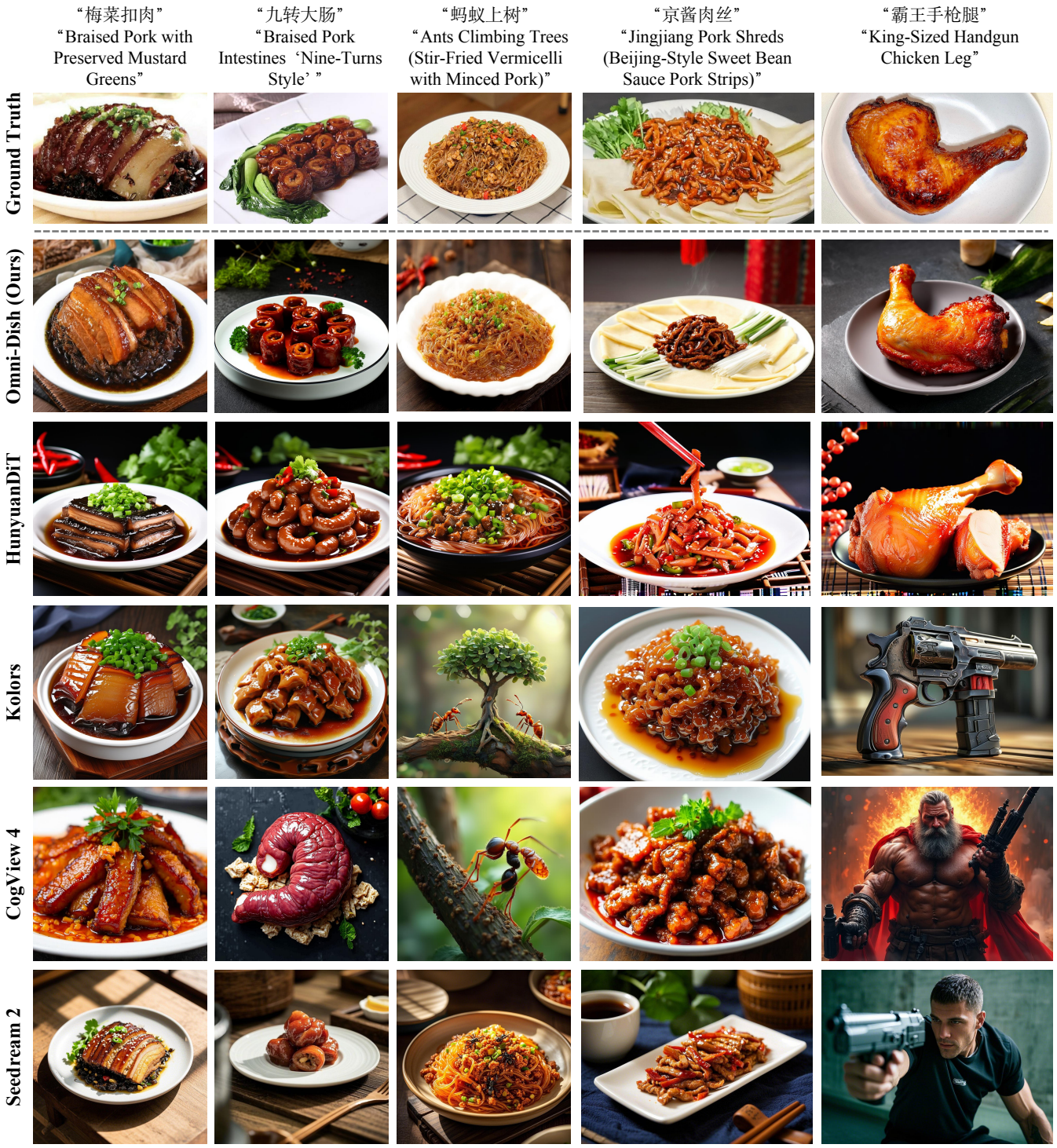

Figure 7: Visual comparison of different models in dish generation. Row 1 displays authentic photos for readers unfamiliar with the dish, with detailed Chinese dish descriptions available in the appendix. Images in the left three columns have a resolution of 1024×1024 pixels, whereas images in the right two columns are 1024×768 pixels.

Table 2: Comparison of Chinese-Compatible T2I models. The introduction to the metrics can be found in Sec. 5.1 and Appendix. We generate 3,000 samples for professional annotation teams to label over 7 days to ensure the credibility of human evaluation.

	Automatic Evaluations		Human Evaluations					
Methods	FID↓	DishSim. ↑	Fidelity ↑	Texture ↑	Composition ↑	Scene ↑	Lighting ↑	Subject ↑
HunyuanDiT [21]	32.85	0.5909	2.027	2.354	2.943	2.524	2.711	2.545
Kolors [44]	31.78	0.5710	2.087	2.834	2.860	2.828	2.972	2.873
CogView4 [46]	27.46	0.5038	1.789	2.747	2.795	2.863	2.854	2.807
Seedream2.0 [14]	19.51	0.6236	2.265	2.905	2.882	2.866	**2.982**	2.901
Omni-Dish(Ours)	**14.96**	**0.7004**	**2.691**	**2.947**	**2.990**	**2.891**	2.920	**2.913**

5 Experiment

5.1 Metrics

Automatic Evaluations. For image generation, existing benchmarks [19, 20, 33] often employ specified prompts to evaluate models in different dimensions, such as numeracy and relationship, which are inadequate for dish generation. Furthermore, since dish images typically exhibit richer visual details and we specifically focus on Chinese dish names, conventional metrics like CLIP T2I similarity [35] and DINO T2I similarity [6] show limited accuracy in measuring the alignment between dish names and images. Thus, we pioneered the use of Dish T2I Similarity, a specialized metric designed to compute the similarity between Chinese dish names and visual content, demonstrating superior performance in handling Chinese dishes. In addition, we constructed a dedicated dataset of 22,000 high-quality dish images to measure FID-22K.

For dish editing, we used CLIP image similarity [35] and DINO [6] image similarity between generated images and ground truth images as evaluation metrics, following [54, 56].

Figure 8: With the "low-refinement" tag, Omni-Dish generates less polished but more realistic and authentic images.

Human Evaluations. We meticulously designed a human evaluation system for dish images. We briefly introduce the meaning of each evaluation dimension below, using a 3-point (1/2/3) scoring system. For dish generation, we evaluated models from 6 dimensions, including fidelity, texture, composition, scene, lighting, and completeness of the subject. For dish editing, we evaluated these methods from 3 dimensions, including effectiveness (Effect.), consistency (Consist.) and aesthetics (Aes.). We provide a detailed introduction to the measurement standards of these metrics in the Appendix.

5.2 Baselines

For dish generation, given the lexical gap in translating culturally Chinese dish names to English, we conducted experiments with four state-of-the-art Chinese-prompt capable generative models: Kolors [44], HunyuanDiT [21], Cogview4 [46], Seedream2.0 [14].

For dish editing, current image editing paradigms predominantly support English instructions. Thus, we compared our model with IP2P [4], HIVE [55], HQ-Edit [23], and MagicBrush [54].

5.3 Evaluation on Text-to-Dish Generation

The quantitative results are presented in Table 2. Omni-Dish achieves state-of-the-art performance across most metrics, particularly in fidelity, demonstrating its capability to generate faithful arbitrary dishes. While its performance in "Lighting" is marginally lower than that of certain methods, it surpasses all other models..

We provide extensive visualizations in Fig. 7. As shown in column 3, "Ants Climbing Trees" is a Chinese dish composed of vermicelli and minced meat. Omni-Dish accurately generated both minced meat and glass noodles with authentic coloration, whereas comparative models misinterpreted the dish's nomenclature or produced unrealistic textural representations. Column 5 of Fig. 7 also shows similar results. In other cases, while existing models could generate visually similar dishes, they failed to preserve faithful nuances. For example, in column 4, "Shredded Pork in Beijing Sauce", typically served with cucumber strips and scallion shreds, other models omitted these essential accompaniments.

In addition, as shown in Fig. 8 with low-refinement tags, Omni-Dish can generate low-refinement but more realistic dish images.

5.4 Evaluation on Dish Editing

In Tab. 3, we compared the performance of several editing methods for dish images across 5 dimensions. Our method significantly outperforms existing models. Although MagicBrush maintains better consistency (Consist.), its editing performance (Effect.) is poor.

We present some visual results in Fig. 9. As shown in the first row ("Remove red peppers from Kung Pao Chicken."), other methods

Figure 9: Visual comparison of different models in dish editing.

Table 3: Quantitative Comparison of Image Editing Methods.

	Automatic Evaluations		Human Evaluations		
Methods	CLIP-I ↑	DINO ↑	Effect. ↑	Consist.↑	Aes. ↑
IP2P [4]	0.8717	0.7554	1.61	1.57	1.43
HIVE [55]	0.8905	0.8096	1.53	1.65	1.76
HQ-Edit [23]	0.7136	0.4541	1.43	1.17	1.31
MagicBrush [54]	0.9437	0.9021	1.71	**2.51**	1.98
UltraEdit [56]	0.9075	0.8433	2.22	1.90	2.02
Ours	**0.9491**	**0.9057**	**2.83**	2.30	**2.44**

fail to recognize what red peppers are and demonstrate nearly zero removal effect, while our method perfectly eliminates all red peppers from the dish. Other examples exhibit similar superior performance.

5.5 Effect of Reception Strategy

We validated the impact of employing a reception strategy on model training. To conserve computational resources, we conducted the experiment with limited training (1M samples). As illustrated in Fig. 10, with the reception strategy, the contours of abalone in "Fo Tiao Qiang" become more distinct, while the outer layer of "Pot-fried pork" clearly exhibits a sweet and sour sauce coating.

6 Conclusion

In this paper, we propose Omni-Dish, the first image generation model specifically designed for Chinese dishes. The generated images exhibit not only photorealism but also faithfully capture the intricate ingredient compositions and cooking details of the dishes. Building upon our generation model, we further develop an editing

Figure 10: Effect of our reception strategy. The reception help the model learn more details, such as the abalone in Fo Tiao Qiang and the sweet and sour sauce in Pot-fried pork.

model to achieve high-quality dish editing. Extensive experiments demonstrate the impressive effectiveness of our approach.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023).

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.

[5] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. InternLM2 Technical Report. arXiv:2403.17297 [cs.CL]

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

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

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

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24185–24198.

[10] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[11] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2023. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102* (2023).

[12] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).

[13] Gabriel Goh, James Betker, Li Jing, and Aditya Ramesh. 2023. *DALL·3*. Retrieved March 26, 2025 from https://openai.com/index/dall-e-3/

[14] Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. 2025. Seedream 2.0: A Native Chinese-English Bilingual Image Generation Foundation Model. *arXiv preprint arXiv:2503.07703* (2025).

[15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

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

[19] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20406–20417.

[20] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023), 78723–78747.

[21] Minbin Huang, Yanxin Long, Xinchi Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu, and Wei Liu. 2024. Dialoggen: Multi-modal interactive dialogue system for multi-turn text-to-image generation. *arXiv preprint arXiv:2403.08857* (2024).

[22] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8362–8371.

[23] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990* (2024).

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

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[27] Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

[28] Sharon Lee, Yunzhi Zhang, Shangzhe Wu, and Jiajun Wu. 2023. Language-informed visual concept learning. *arXiv preprint arXiv:2312.03587* (2023).

[29] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10758–10768.

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[31] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. 2025. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487* (2025).

[32] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

[33] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855* (2024).

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2023), 53728–53741.

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[38] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159 [cs.CV]

[39] Elad Richardson, Yuval Alaluf, Ali Mahdavi-Amiri, and Daniel Cohen-Or. 2024. pOps: Photo-inspired diffusion operators. *arXiv preprint arXiv:2406.01300* (2024).

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[42] 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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[43] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2024. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8871–8879.

[44] Kolors Team. 2024. Kolors: Effective Training of Diffusion Model for Photorealistic Text-to-Image Synthesis. *arXiv preprint* (2024).

[45] Qwen Team. 2025. Qwen2.5-VL. https://qwenlm.github.io/blog/qwen2.5-vl/

[46] Zhipu AI Tsinghua University. 2025. *CogView4 and CogView3 and CogView-3Plus*. Retrieved March 21, 2025 from https://github.com/THUDM/CogView4

[47] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. 2023. Semi-supervised parametric real-world image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5927–5936.

[48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* (2024).

[49] Zhiling Wang, Weiqing Min, Zhuo Li, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2022. Ingredient-Guided Region Discovery and Relationship Modeling for Food Category-Ingredient Prediction. *IEEE Transactions on Image Processing* 31 (2022), 5214–5226. doi:10.1109/TIP.2022.3193763

[50] Zhiling Wang, Weiqing Min, Zhuo Li, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2022. Ingredient-guided region discovery and relationship modeling for food category-ingredient prediction. *IEEE Transactions on Image Processing* 31 (2022), 5214–5226.

[51] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).

[52] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[53] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024. AnyEdit: Mastering Unified High-Quality Image Editing for Any Idea. *arXiv preprint arXiv:2411.15738* (2024).

[54] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems* 36 (2023), 31428–31449.

[55] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. 2024. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9026–9036.

[56] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2024. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. *Advances in Neural Information Processing Systems* 37 (2024), 3058–3093.

[57] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *European Conference on Computer Vision*. Springer, 1–22.