X2I: Seamless Integration of Multimodal Understanding into Diffusion Transformer via Attention Distillation

Jian Ma ✉
OPPO AI Center
majian2@oppo.com

Qirong Peng*
OPPO AI Center
pengjirong@oppo.com

Xu Guo†
Tsinghua University
guo-x24@mails.tsinghua.edu.cn

Chen Chen ✉
OPPO AI Center
chenchen4@oppo.com

Haonan Lu
OPPO AI Center
luhaonan@oppo.com

Zhenyu Yang
OPPO AI Center
yangzhenyu@oppo.com

Abstract

Text-to-image (T2I) models are well known for their ability to produce highly realistic images, while multimodal large language models (MLLMs) are renowned for their proficiency in understanding and integrating multiple modalities. However, currently there is no straightforward and efficient framework to transfer the multimodal comprehension abilities of MLLMs to T2I models to enable them to understand multimodal inputs. In this paper, we propose the X2I framework, which endows Diffusion Transformer (DiT) models with the capability to comprehend various modalities, including multilingual text, screenshot documents, images, videos, and audio. X2I is trained on a 100K English corpus in 160 GPU hours. Building on the DiT teacher model, we adopt an innovative distillation method to extract the inference capabilities of the teacher model and design a lightweight AlignNet structure to serve as an intermediate bridge. Compared to the teacher model, X2I shows a decrease in performance degradation of less than 1% while gaining various multimodal understanding abilities, including multilingual to image, image to image, image-text to image, video to image, audio to image, and utilizing creative fusion to enhance imagery. Furthermore, it is applicable for LoRA training in the context of image-text to image generation, filling a void in the industry in this area. We further design a simple LightControl to enhance the fidelity of instructional image editing. Finally, extensive experiments demonstrate the effectiveness, efficiency, multifunctionality, and transferability of our X2I. The open-source code and checkpoints for X2I can be found at the following link: https://github.com/OPPO-Mente-Lab/X2I.

*Co-first authors
†The author did his work during internship at OPPO AI Center.
✉Corresponding authors.

1. Introduction

The recently open-sourced T2I models[50, 54, 61, 62, 66], such as Flux.1[32], have ushered in a new era of AI art due to their ability to generate realistic, photograde images that are both interesting and creative. The framework of T2I models has evolved from early GAN-based models[16, 27], auto-regressive models[60, 85], and UNet-based diffusion models[54] to DiT models[8, 17, 71]. To enhance both controllability and practicality in visual generation, numerous implicit[41, 46, 48] and explicit[34, 47, 84, 88] instruction-based editing models leveraging reference images have been developed quickly. Additionally, several unified image instruction editing models[22, 68, 86] have appeared. They all share a critical characteristic: the requirement of collecting comprehensive instruction-based editing datasets and committing to costly training resources.

The text encoder in T2I is critical for generating semantically relevant images, as seen from early models such as CLIP[57] to later models such as T5[58] and advanced LLMs[13, 72]. The main goal is to improve the semantic understanding of text without considering the impact of other modalities on visual output. In contrast, VLMs, including early approaches like Flamingo[2] and BLIP2[35], along with newer models like Llama-Vision[12], QwenVL[3], and InternVL[10], have effectively integrated pre-trained visual encoders with LLMs for enhanced text and visual comprehension. Is there a more streamlined way to transfer VLM or MLLM capabilities to DiT to enable diverse modal inputs? The industry is primarily tackling this challenge through two strategies. One method, similar to PEA-Diffusion[45], utilizes feature distillation by aligning multilingual LLM encoders to the original dimensions using MLP. However, it requires substantial resources of 1600 GPU hours and large-scale image-text datasets. This distillation relies on the U-Net[63] block's

Figure 1. The primary applications of X2I include multilingual text-to-image, image-to-image, image-text-to-image, video-to-image, audio-to-image, and various multimodal combinations for image generation. The red boxes embedded in the images indicate the input images, text, video, or audio. The red box text below the images represents the abbreviated application names, with the initial letters of different modalities in parentheses. The descriptions below some of the images provide a brief overview of the input audio or video.

outputs and restricts support for multimodal input. Another method is the VLMs-based replacement of T2I[43], but it is costly, requiring extensive training or fine-tuning of T2I or MLLMs, necessitating substantial data quality and quantity[22, 68]. Alternatively, a solution akin to GlueGen[55] could be employed to align the features of a single-modal or multimodal encoder with the latent space of existing T2I models. However, aligning solely through the encoder still results in suboptimal overall performance.

In this work, we introduce an efficient framework, X2I, along with a simple LightControl module to facilitate the enhancement of image transition from weak to strong fidelity. Utilizing a DiT-based image generation model as a teacher, the student model mirrors the entire architecture of the teacher model but substitutes its text encoder with MLLMs. Both models are capable of processing the same linguistic input. To optimize our training data without gathering images, we opt to distill the teacher model's inference abilities. In T2I models, this necessitates performing distillation during the reverse denoising phase. Furthermore, we efficiently transfer the visual generation capabilities of the teacher model to the student model by ingeniously designing the AlignNet structure and distilling the attention. As a result, after alignment, the student model acquires the multimodal comprehension abilities of MLLMs during image generation. Moreover, we formulate a template input format intended for the student model to effectively interpret instructions from different modalities. Since MLLMs generally have proficiency in understanding the global semantic features of visual content, we develop LightControl to derive structured data from reference images. This enhances visual fidelity and facilitates precise and controllable image editing.

After training the X2I, we conduct various experiments to compare the capabilities of T2I and instructional-based image editing. Additionally, we conduct subjective experiments for image generation with various multimodal mixed inputs, including multilingual, image (I2I), image-text (IT2I), video (V2I), and audio (A2I). The results demonstrate that X2I acquires robust multimodal understanding capabilities with less than 1% performance degradation in the T2I generation. Furthermore, for more personalized tasks like I2I or IT2I, we conduct experiments involving the training of traditional LoRA[23] to fulfill in-

dustry requirements, addressing a training gap in LoRA for these purposes. Moreover, to enhance the adaptability of the framework, only the AlignNet parameters within the model are updated, enabling the aligned student model to accommodate a variety of downstream tasks, such as LoRA, ControlNet[70], IP-Adapter[84], different fine-tuning models, and compression models. Notably, some objective metrics for tasks involving ControlNet and IP-Adapter show improvements over the teacher model. This improvement is due to X2I's robust support for image conditional information input, which facilitates further extraction of image features. We also perform experimental analyses on training efficiency, attaining 98.2% of the teacher model's performance with 48K training samples.

In summary, our key contributions are as follows.

- We introduce a novel method for transferring the understanding capabilities of MLLMs to DiT models, achieving fast fitting speeds with only a small amount of monolingual text data through the carefully designed AlignNet and Attention Distillation.
- We design LightControl to enable the generation of image instruction editing from weak fidelity to high fidelity.
- The architecture itself possesses strong modularity, leading to performance improvements in certain IT2I tasks. It also has the ability to train LoRA in IT2I tasks.
- X2I is the first image generation model that supports audio understanding in addition to text and visual understanding.

2. Related Work

Diffusion Models and Text Encoder. Diffusion models have shown immense potential in the field of T2I. Earlier works[17, 62, 65, 82] use U-Net as the backbone of the diffusion model, which, to some extent, impacts the scalability of the models. Inspired by DiT[52], many works[8, 17, 32, 36] replace the U-Net backbone with transformers, greatly improving the quality of the generated images. In order to improve the understanding of the diffusion model's prompts, Stable Diffusion (SD) used CLIP to encode text information, while Flux.1 and eDiff-I[5] utilize information encoded by both T5 and CLIP to guide image generation. Recently, works like LuminaT2X[19], Kolors[71], and LI-DiT[44] directly use pre-trained decoder-only LLMs as the prompt encoder, which enhances the prompt-following ability in image generation. However, these methods can only encode the text prompt, lacking the ability to understand information from images, videos, audio, etc.

MLLMs as Text Encoders for T2I. VLMs[3, 10, 35, 39, 42, 74, 75, 78] have made significant progress in visual language understanding tasks. Furthermore, models such as AnyGPT[87], NextGPT[76], X-LLM[7], and MiniCPM-o[24] have further developed the capabilities in audio understanding. In articles that fuse multimodal models with image generation models, MUMU[6] and ML-MGIE[18] inject multimodal information into the image generation model using VLM but also change the weights of the generation model during training, lacking plug-and-play capabilities. KOSMOS-G[51] and Easyref[91] achieve alignment while preserving the weights of the original SD[16] and SDXL[54] models. KOSMOS-G trains a VLM and AlignerNet from scratch, whereas Easyref trains the features and projector of the final layer output of the VLM. These methods consume significant computational resources and are limited to a single task, failing to fully leverage the zero-shot capabilities of VLM. Furthermore, the aforementioned methods focus on alignment with U-Net-based models like SDXL. Currently, there is a lack of an efficient alignment method for DiT models, such as Flux.1 and SD3. Work in the related field, such as HunyuanVideo[30], requires an additional CLIP text encoder for alignment. Qwen2VL-Flux[43] not only retains the structure of T5 but also alters the weights of Flux.1 itself, significantly reducing the model's plug-and-play and zero-shot capabilities.

Distillation of DiT. Knowledge distillation (KD) has been widely applied in diffusion models[28, 89, 90]. SSKD[73] involves the direct distillation of critical attention mechanism information from teacher to student, which can considerably reduce the performance gap between both. EFFICIENT-VDIT[14] utilizes a multi-step consistency distillation technique to accelerate DiT sampling. PEA-Diffusion integrates multilingual CLIP with SDXL by employing the L2-norm distance, thus allowing the T2I model to support multilingual capabilities. However, this method is specifically tailored for the U-Net backbone and demands significant computational resources.

3. Methodology

3.1. Preliminary

The diffusion process is performed in the latent space, where a transformer denoiser ϵ_θ is employed to predict noise ϵ with the current timestep t, noisy latent x and generation conditions, c and c_p, where $c_p = \tau_\theta(y)$ is produced by encoding the text prompts y with a pre-trained CLIP text encoder τ_θ, and $c = \varepsilon_\theta(y)$ is produced by encoding the text prompts y with a pre-trained T5 text encoder ε_θ. To enhance the capability of feature fusion, MM-DiT[17] uses adaptive layer normalization (AdaLN) to improve the adaptability of the model when processing different input conditions. There are two fully connected layers plus a SiLU activation function forming the feature extraction and normalization networks ζ_θ and δ_θ , which respectively extract features for x and c. These features are used as the normalization and scaling parameters for the multi-head self-

attention, as shown below:

$$x, \alpha_1, \beta_2, \gamma_2, \alpha_2 = \zeta_\theta(x, t, c_p)$$
$$c, \alpha_{1c}, \beta_{2c}, \gamma_{2c}, \alpha_{2c} = \delta_\theta(c, t, c_p), \quad (1)$$

then, concatenate x and c directly with $\varphi(\cdot)$ and use the self-attention to achieve feature interaction and fusion. The output is then split according to the corresponding indices, as shown below:

$$x_A, c_A = softmax\left(\frac{QK^T}{\sqrt{d_0}}\right) \cdot V, \quad (2)$$

where $Q = W_Q \cdot \varphi(x, c), K = W_K \cdot \varphi(x, c), V = W_V \cdot \varphi(x, c)$, x_A, c_A represents the feature value output on the image and text sides obtained after computing attention and truncating according to the respective index values. W_Q, W_K, W_V are learnable projection matrices.

Both the text and image sides undergo the following process, and only the image-side processing is described below for the convenience of presentation. After the residual module, the regressed scaling parameter α_1 is used to scale the weights of the image-side input,

$$x_{LN} = \mu(x + \alpha_1 * x_A), \quad (3)$$

where μ represents Layer Normalization (LN). Then, normalization is performed utilizing adaptive parameters β_2, γ_2,

$$x_{FF} = \nu(x_{LN} * (1 + \gamma_2) + \beta_2), \quad (4)$$

where ν represents FeedForward (FF). Finally, the output of the current block is obtained by multiplying the regression scaling coefficients and the residuals

$$x_O = (x + \alpha_1 * x_A) + x_{FF} * \alpha_2. \quad (5)$$

3.2. Model Overview

The framework consists of two parts, as shown in Fig. 2. The initial part involves X2I training exclusively on a text corpus, with AlignNet as the trainable parameters. Distribution alignment between student and teacher is attained via attention distillation. The subsequent part incorporates the LightControl module, forming X2I Enhanced Training, thereby improving the accuracy of reference images in image instruction editing tasks.

3.3. AlignNet with MLLM's Hidden State

Many approaches use MLLM as text encoders for T2I or feature extractors for other modules, typically utilizing only the last[51] or penultimate[79] hidden states. However, according to Oscar Skean's analysis[69], the quality of hidden states in the intermediate layers of transformer-based LLMs is often superior to that of the final and initial layers. Recognizing that employing a static intermediate layer's

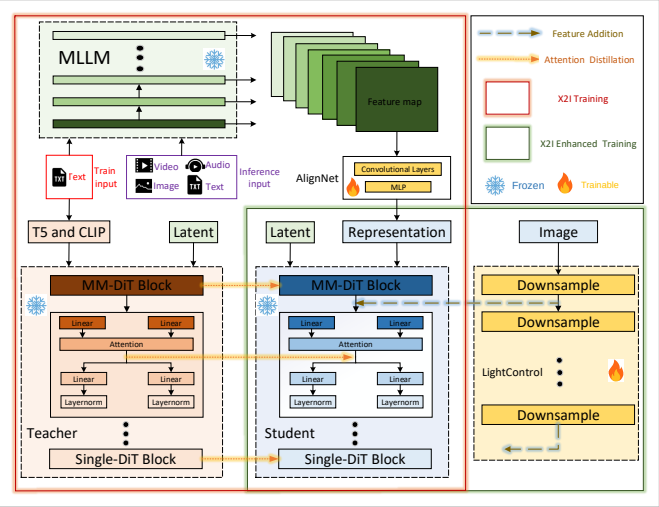


Figure 2. Overview diagram of the X2I. In the deep red box on the left represents the X2I training, where only the text data needs to be input for training, with trainable parameters limited to AlignNet and the distillation location situated within the attention output of MM-DiT. The light green box at the bottom right represents X2I enhanced training after the X2I training, AlignNet is fixed, and the trainable parameters limited to LightControl.

representation might be suboptimal for adapting to diverse MLLM models, we propose utilizing representations from every layer of MLLMs as input to AlignNet.

To integrate hidden states from all layers of MLLMs, we design a simple CNN to capture spatial and channel-wise relationships between layers. Convolutional operations allow the extraction of more complex patterns and improve the rationality of weight allocation, enabling spatial adaptability and establishing inter-layer dependencies. By adjusting the kernel size, we can also create cross-layer receptive field correlations, achieving better multi-scale fusion.

In MLLM, define a feature extractor that projects text y into a middle feature $H \in \mathbb{H}^{b \times m \times s \times z}$, where m represents all layers of MLLM's hidden states, each layer has a hidden size of z and a sequence length of s. Next, we define a CNN mapping network Ψ to enable self-learning of weights across all layers' hidden states, resulting in the aggregation of hidden states from m layers into 1 layer of weighted hidden states as the following formula,

$$y_p, y = \Phi(\Psi(m, 1, k, p)(H)), \quad (6)$$

where the kernel size is k, and the padding is denoted as p, Φ is a simple MLP network that maps the feature learned by the Ψ with weight information to the corresponding dimension of c and c_p, thereby obtaining new features y and y_p.

3.4. Attention Distillation

KD can be divided into two main categories: logits distillation and feature distillation. In MM-DiT, intermediate layer features encompass low-level information such as pixel values, shape, and gradient information, as well as high-level information such as color, theme, and lighting details. In this paper, the student model is conditioned to inject multimodal feature information. Knowledge transfer occurs layer by layer through the concatenation of the information injected at each level. Therefore, for deeply encoded information that logits cannot express, we opt for intermediate feature distillation between layers to achieve alignment.

Further, each MM-DiT consists of AdaLN as Eq. (1), self-attention layer as Eq. (2), LN as Eq. (3), and FF layers as Eq. (4). In addition to regressing the γ and β after the LN, AdaLN also regresses an α before the end of each residual module. Since biases here directly affect the distribution of features, we choose to perform feature distillation in the position of the attention output x_A, c_A in order to minimize unnecessary regression and maintain consistency in the distribution affected by normalization parameters. This provides an important conclusion in exploring direct feature distillation in the MM-DiT feature layers as well as the the single-DiT block of Flux.1.

Regarding the choice of attention distillation loss, we adopt reverse KL (RKL) from MiniLLM[20]. KL performs well in traditional tasks due to the smaller output space and fewer modes in conventional classification tasks. However, for LLMs, the output space is more complex with more modalities; RKL can prevent the student model from overfitting to the low-probability outputs of the teacher model, and the same applies to the MM-DiT structure. For the Flux.1 structure, the target loss for l layers of MM-DiT is defined as follows:

$$\mathcal{L}(\theta) = \mathrm{KL}[\phi_l || \upsilon_l] = \left[-\mathbb{E} \log \frac{\upsilon_l(A_t|\epsilon, c)}{\phi_l(A_s|\epsilon, y)} \right], \quad (7)$$

where ϕ, υ specifically represent the attention output mappings in each layer of the student and teacher models in MM-DiT, respectively. A_t denotes the attention output x_A, c_A mentioned above; similarly, A_s corresponds to the output of the attention output in the student model. ϵ represents noise input because we perform distillation during the inference process, which is the reverse denoising process.

3.5. Training Stages

For most pre-trained MLLMs, features from different modalities are mapped to a common feature space, where similar information from different modalities is brought closer together. This implies that different modality features share a common semantic space. Based on this consensus, we opt to perform alignment by inputting pure text modality while simultaneously aligning the inferential capabilities of the teacher model, that is, aligning capabilities during the process of pure noise denoising. The Appendix Sec. 8 Fig. 7 provides a more detailed analysis of the changes in the distribution of different modality spaces before and after alignment, which effectively explains how X2I gained the multimodal understanding capabilities of MLLM simply through textual semantic alignment.

Since MLLMs tend to understand images more from a global semantic perspective, they often lack the crucial fine-grained understanding needed for image generation. To address this issue, we design a simple LightControl in parallel with the MM-DiT structure named X2I enhanced training. This pathway is a convolutional module composed of ResNet blocks, similar to those mentioned in ControlNeXt[53]. We initialize a ResNet network with 19 layers, mirroring those used in MM-DiT. We define the ResNet module Γ to extract the conditional controls; similar to ControlNet, the output of each ResNet module is:

$$y_c^l = \Gamma^l(c_p, c_i; \Theta), \quad (8)$$

where y_c^l is the feature output of l the layer of the ResNet, c_p is the corresponding conditional text input, c_i is the conditional reference image input, and Θ is the corresponding trainable parameter. The final output here is also similar to ControlNet, where y_c and x_O are added feature-wise.

4. Experiments

4.1. Datasets and Implementation

We randomly select 100K image-text pairs from the Laion2B [67], and utilize InternVL to generate new captions from the images for training X2I. Furthermore, based on the T2I-Adapter's[49] sketch-guided model and SDXL, we construct a training dataset of 400k images containing five styles: Monet, Baroque, Cartoon, Pixar, and Van Gogh. For the text encoder in the student model, we utilize open-source models from the InternVL series, Qwen2.5-VL-7B[4], and MiniCPM-o, for the DiT structure, we use Flux.1. To optimize performance, inference ability is distilled by focusing on the denoising process between steps 0 and 1. The X2I is trained on 8 A100 GPUs for 24K steps. X2I training requires over 80GB of GPU memory. We optimize this by separating training and inference, and overlapping communication, achieving a global batch size of 6. Acceleration techniques are detailed in the appendix Sec. 9.

4.2. Tasks and Benchmarks

T2I and Multilingual T2I. **Evaluation Datasets**. We randomly select five sub-evaluation datasets from T2I-CompBench++[25], each containing 1k images. These include: complex compositions bench, non-spatial bench, attributes binding bench related to color and texture categories, numeracy bench. GenAI-Bench[33], EvalMuse[21],

and Multilingual-General[45]. **Metrics** includes four main categories: image-text matching score indicators like ClipScore (CS) with ViT-bigG[11] and FGA-BLIP2 (FB)[21] Score. Human-consistency feedback score indicators include PickScore (PS)[29], HPSv2[77], and ImageReward (IR)[81]. Fine-grained score indicators have BLIP-VQA (BV)[25] and UniDet[25]. Model scoring indicators reflecting complex textual semantic alignment such as GPT-4o and VQAScore (VS)[38], with GPT-4o instructions specified by the T2I-CompBench++ for complex compositions evaluation. Furthermore, the IR score ranges from 98.2% probability within [-2, 2], and the FB score ranges from [1, 5]. We use eight normalized metrics to determine the relative gap compared to the teacher model, defining a Performance Ratio (PR) metric as a performance percentage indicator. **Baseline model** for universal T2I capability comparison is the Flux.1 teacher model, used to assess reduction in student model's basic T2I capabilities. Qwen2VL-Flux has similar functions, but MLLM does not support text input and relies solely on the source model's T5. Thus, it is excluded from comparison. For the comparison of multilingual T2I capabilities, we evaluate both PEA-Diffusion and Sana[80].

IT2I and I2I. **Evaluation Datasets**. We chose DreamBench[64] and we generate six images for each text prompt[34]. **Metrics**. We report the average DINO, CLIP-I, and CLIP-T scores based on all pairs of real and generated images. **Baseline model**. The comparison model includes methods based on different frameworks such as SD, Imagen, Phi-3, SD3, and Flux.1. In the appendix Sec. 12, we also compare the differences in Qwen2VL-Flux's capabilities in image understanding and image editing based on objective metrics.

Image stylization. **Evaluation Datasets** consists of 50 images from MSCOCO[37] and 50 photorealistic images generated by DiffArtist[26]. **Baseline model** is also selected DiffArtist.

ControlNet. **Evaluation Datasets** is sourced from MultiGen-20M[56] and includes four tasks: A model based on canny edge detection, evaluated using the F1 score. Models based on hed, evaluated using Structure Similarity Index Measure (SSIM). A model based on depth information is evaluated using RMSE.

4.3. Quantitative Results

Universal T2I Generation. As shown in Tab. 1, the one that performs closest to the teacher model is the complex compositions evaluation set. Additionally, on some sub-evaluation sets and certain metrics, the performance of X2I even surpass that of the Flux.1. The overall average PR across the five evaluation sets is 99.21% of that of Flux.1, with such subtle differences being virtually indistinguishable subjectively. Appendix Sec. 10 Tab. 4 displays the results from the GenAI-Bench and EvalMuse evaluation sets, with performance scores also exceeding 99%. This further demonstrates that X2I's results in T2I generation are nearly indistinguishable from those of the teacher model Flux.1.

Multilingual T2I Generation Tab. 5 includes three multilingual evaluation datasets from Multilingual-General, as well as four additional translations in German, Portuguese, Spanish, and French. While the performance in English remains consistent with a one percentage point difference similar to before. For more details, please refer to Sec. 11.

IT2I Generation Tab. 2 shows the performance of different methods on the DreamBench. Overall, X2I has a slight advantage in DINO scores. CLIP-I shows a two-point difference compared to KOSMOS-G and there is a considerable gap in CLIP-T compared to recent methods like OmniControl and IP-Adapter based on Flux.1, indicating that X2I still has room for improvement in its ability to faithfully follow the semantics of images.

4.4. Qualitative Results

X2I training achieves outstanding results through feeding pure text. Fig. 1 illustrates the effects of some of these functions. T2I possesses the capability to support over 29 natural languages, covering the major language families worldwide; I2I includes image variations and the fusion of concepts from multiple images. IT2I encompasses single image instruction editing, style transfer, image stylization, instruction editing for multiple images, and the result of text token mixing with multiple images. It integrates multiple visual concepts to create unique images, extracts specific concepts from a set of input images, and synthesizes these concepts into new images, allowing precise control over the visual details of the images and avoiding common vagueness and uncertainty in text descriptions. Examples include changing clothes in the combination of clothes and people, performing group photos in person combinations, and creative conceptual combinations with three or more images.

A2I enables auditory information to be visualized, providing a way to "see sound". This aspect is not limited to regular audio information but also processing various audio types, including natural water flow sounds, animal calls, iconic IP character voices, and the emotions that different music pieces convey, translating these into corresponding visual images. V2I can inductively summarize single frames from videos as images and generate high-definition images from low-resolution videos.

Additionally, more creative combinations can be achieved. In the middle of Fig. 1, the last three examples in the second row illustrate this: the first input example is a 1-minute video generated by Sora[40], a sunset background image along with the text instruction "wearing a hat and scarf", the generated image successfully maintains the original video's semantic information while adhering well

T2I++ CompBench	Methods	CS	PS	HPSv2	IR	FB	GPT-4o	VS	BV	UniDet	PR
Complex	Flux.1	0.4279	0.2297	0.2055	0.8354	3.6990	0.9181	0.8799	0.5402	-	99.77%
	X2I	0.4234	0.2283	**0.2096**	0.8124	3.6723	0.9142	0.8779	**0.5417**	-	
Non-spatial	Flux.1	0.4418	0.2751	0.2472	1.3013	3.5963	0.8962	0.9075	0.6762	-	99.12%
	X2I	0.4363	0.2706	0.2462	1.2786	3.5449	0.8821	0.8996	0.6571	-	
Color	Flux.1	0.4733	0.2872	0.2417	1.4179	4.2268	0.7060	0.9077	0.6642	-	99.72%
	X2I	0.4691	**0.2881**	**0.2421**	1.3872	4.1845	**0.7113**	0.9014	**0.6643**	-	
Texture	Flux.1	0.4572	0.2589	0.2411	1.1994	4.0601	0.7588	0.8807	0.5281	-	99.18%
	X2I	0.4504	0.2570	0.2401	1.1287	4.0055	0.7447	**0.8809**	0.5172	-	
Numeracy	Flux.1	0.4730	0.2664	0.2431	1.3459	3.4852	0.7860	0.7841	0.4528	0.6112	99.36%
	X2I	0.4713	**0.2670**	**0.2485**	1.2723	3.4140	0.7680	0.7798	0.4464	**0.6206**	

Table 1. Objective performance on the T2I-CompBench++ evaluation dataset.

Methods	Base	DINO	CLIP-I	CLIP-T
Textual Inversion[83]		0.569	0.780	0.255
DreamBooth[64]		0.668	0.803	0.305
Custom Diffusion[31]		0.643	0.790	0.305
BLIP-Diffusion[34]	SD[62]	0.594	0.779	0.300
Subject-Diffusion[47]		0.711	0.787	0.293
IP-Adapter[84]		0.667	0.813	0.289
KOSMOS-G[51]		0.694	**0.847**	0.287
SuTI[9]	Imagen[65]	0.741	0.819	0.304
OmniGen[51]	Phi-3[1]	-	0.801	0.315
UNIC-Adapter[15]	SD3[17]	0.816	0.841	0.306
IP-Adapter[84]		0.768	0.803	0.322
OminiControl[70]	Flux.1[32]	0.740	0.768	**0.329**
X2I		**0.817**	0.826	0.304

Table 2. Performance of different methods on DreamBench.

to both the text and image instructions. The second input combines an image of the Sanxingdui bronze sacred tree, a sci-fi novel text fragment, and a sound of mechanical operation, resulting in an output that blends all the semantic information, adding a mysterious touch. The third input is a Chinese ink painting, a classical poem, and Beethoven's Moonlight Sonata, resulting in a final image that perfectly integrates the three modalities into an evocative artwork. For more interesting effect displays, please refer to the appendix Sec. 17.

X2I enhanced Training. The purpose of constructing LightControl is to enhance the fidelity of image instruction editing. Essentially, it is not limited to a specific downstream task. In this paper, relevant experimental validation is only conducted in the stylization domain. For more comparative results, please refer to the appendix Sec. 16.

IT2I LoRA. For highly personalized IT2I image generation tasks tailored to the user, we can leverage the X2I to train a LoRA. In this paper, we train X2I using reference images in the style of abstract art line drawings. More comparative results, please refer to the appendix Sec. 15.

Plug-and-Play. In this paper, we have validated the prominent variant models and related downstream models in the Flux.1 community. Specifically, we compare objective metrics for tasks involving ControlNet and IP-Adapter. Additionally, incorporating the original image as input in the X2I task resulted in improvements in both objective and subjective metrics. More results in the appendix Sec. 13.

Reasoning ability and multi-turn dialogue generation ability MLLMs themselves possess certain reasoning abilities and multi-turn dialogue capabilities. X2I also inherits some of these related abilities. More results in the appendix **??**.

4.5. X2I Training Convergence

Benefitting from the design of AlignNet and attention distillation, the training convergence speed of X2I is remarkably fast. Here, we have calculated the performance metrics including CS, FB Score, PS, IR, and VS, and used the average of these five objective metrics as the Performance Ratio. Additionally, we have also calculated the variation of the SSIM metric. As shown in Fig. 3 with Qwen2.5-VL-7B, the convergence speed in the early stages of training exhibits an almost linear growth, achieving a Performance Ratio of 98.2% with only 8K steps of training. Appendix Sec. 7 further demonstrates the experimental results of InternVL[10] with different capacity adaptations for X2I.

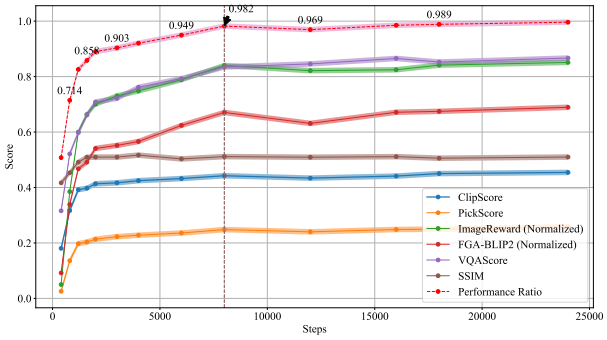


Figure 3. X2I training performance vs. training steps.

4.6. Ablation Study

In this paper, we design three types of ablation experiments: AlignNet structure with the number of feature layers extracted by MLLMs, the feature distillation position in the

Models	CS	PS	IR	FB	SSIM	PR
FLUX.1	0.4517	0.2578	1.3891	3.7789	-	100%
A1	0.4313	0.2339	1.1149	3.4946	0.5089	95.40%
A3	0.4348	0.2407	1.2198	3.544	0.5131	96.62%
A3_ada	0.4429	0.2447	1.2278	3.6055	0.5131	97.36%
A3_t5_ada	0.4397	0.2452	1.2757	3.6016	0.5036	97.57%
A29_ada	0.4463	0.2478	1.2954	3.6789	0.5271	98.40%
A29_t5_ada	0.4361	0.2421	1.1988	3.5712	0.5084	96.73%
A29_CNN	**0.4508**	**0.2518**	**1.3799**	**3.6748**	**0.5189**	**99.11%**
Block	0.1483	-0.0220	-2.2800	1.0854	0.0475	45.65%
FF	0.2995	0.1027	-0.7019	2.2326	0.4623	69.58%
LN	0.4392	0.2495	1.3004	3.6557	0.5128	98.16%
Oneside	0.4470	0.2461	1.3030	3.6312	0.5077	98.13%
KL	0.4528	0.2532	1.4039	3.6988	0.5152	99.26%
JS	0.4490	0.2537	1.3795	3.6972	0.5202	99.26%
RKL	**0.4543**	**0.2577**	**1.4043**	**3.7568**	**0.5208**	**99.70%**

Table 3. Three ablation results of X2I.

MM-DiT structure, and different target losses for feature distillation. The evaluation dataset uses the Multilingual-General. The evaluation metrics are CS, FB Score, PS and IR, and the SSIM metric for image variants. Additionally, the overall PR of these four normalized metrics. The MLLMs chosen is Qwen2.5-VL.

AlignNet Structure Ablation. Tab. 3 upper portion show AlignNet structure ablation. "A1" indicates extracting only the last layer features of Qwen2.5-VL while the AlignNet structure consists of a simple MLP, similar to PEA-Diffusion, performing only simple dimensional mapping. The overall performance of the relevant metrics for T2I generation is 95.4% of the teacher, showing a relatively large decline. "A3" extracts features from the first and the last two layers of Qwen2.5-VL, where the first layer features are static token-mapped vector features trained by MLLMs, containing the most original token information, but lacking contextual information. Therefore, combining with the last two layers features offers complementary advantages. By directly taking the average pooling of the features, the performance improvement is about one percentage point compared to "A1". "A3_ada" further learns adaptive weights for each layer of features, improving by 0.74%. "A3_t5_ada" increases the nonlinearity complexity of AlignNet by adding a custom T5 module for feature mapping, but the improvement is minimal. "A29_ada" selects features from all 29 layers of Qwen2.5-VL, leading to about a one percentage point improvement over "A3_ada". However, "A29_t5_ada" shows some metric decline, suggesting that deep nonlinear complex mapping is not required for such modal alignment learning. Furthermore, the scalar parameter learning fails to capture the synergy between low-level edge features and high-level semantic features, as well as issues of spatial insensitivity and poor dynamic adaptability. Therefore, "A29_CNN" ultimately adopts a CNN to further enhance the non-linear learning of each layer's feature weights, resulting in optimal performance, reducing the performance loss compared to the teacher to less than one percentage point.

Distillation Position Ablation. The middle four lines of Tab. 3 show distillation position ablation. "Block" refers to direct distillation of each layer's MM-DiT and Single-DiT output positions in Flux.1, analogous to PEA-Diffusion's method of distilling each U-Net block output in SDXL. The experimental result shows a suboptimal performance of 45.65%. "FF" denotes distillation at the output position of each block's FeedForward layer, as shown in Eq. (4), results indicate a significant improvement compared to "Block" distillation. This is likely due to the influence of the regression scaling factor α_2 on distribution alignment in Eq. (5). LN denotes distillation at the second layer norm output position of each block, i.e., the output of Eq. (3), approaching optimal results but still possibly affected by the regression scaling factor α_2. "Oneside" indicates distillation solely at the conditional side output after MM-DiT's self-attention, aiming to explore whether the interaction information's purest form can be obtained after self-attention and token length segmentation. Experimental results remain about one percentage point below the optimal results.

Feature Alignment Loss Ablation. We additionally compare three other target loss functions: KL, JS, and RKL divergence. Both KL divergence and JS divergence consistently exhibited slightly better performance compared to the MSE target loss like "A29_CNN" before. However, the best performance was achieved with RKL, which outperformed KL divergence by 0.44% percentage points.

5. Conclusion and Limitations

In this paper, we introduce a novel framework X2I for aligning MLLMs and the DiT visual generation model. This framework can be fully trained with minimal training resources and minimal training data. Our extensive experimental results demonstrate that X2I maintains its general performance while extending additional functionalities, including multilingual image generation, image instruction editing, conceptual fusion of multiple images, image generation from videos or audio, and further exploration of creative combinations from different modalities to unleash imagery with X2I. Additionally, X2I supports LoRA training for multiple modalities, enabling seamless integration with various controllable downstream tasks in the open-source community. Furthermore, we have developed LightControl to explore more precise and controllable possibilities for visual generation.

The limitations of this paper lie in the somewhat inferior precision and control of X2I in the unified direction of image instruction editing, as well as the failure to fully explore the logical reasoning, few-shot learning, and chain-of-thought capabilities of MLLMs. In the future, we will continue to delve deeper into this direction.

References

[1] Marah Abdin and et al Jyoti Aneja, Hany Awadalla. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 1

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1, 3

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 5

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. 3

[6] William Berman and Alexander Peysakhovich. Mumu: Bootstrapping multimodal image generation from text-to-image data, 2024. 3

[7] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages, 2023. 3

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 1, 3

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning, 2023. 7

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1, 3, 7

[11] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022. 6

[12] Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama backbone for vision tasks, 2024. 1

[13] DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. 1

[14] Hangliang Ding, Dacheng Li, Runlong Su, Peiyuan Zhang, Zhijie Deng, Ion Stoica, and Hao Zhang. Efficient-vdit: Efficient video diffusion transformers with attention tile, 2025. 3

[15] Lunhao Duan, Shanshan Zhao, Wenjun Yan, Yinglun Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Mingming Gong, and Gui-Song Xia. Unic-adapter: Unified image-instruction adapter with multi-modal transformer for image generation, 2024. 7

[16] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. 1, 3

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 1, 3, 7

[18] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models, 2024. 3

[19] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, Renrui Zhang, Junlin Xi, Wenqi Shao, Zhengkai Jiang, Tianshuo Yang, Weicai Ye, He Tong, Jingwen He, Yu Qiao, and Hongsheng Li. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers, 2024. 3

[20] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models, 2024. 5

[21] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation, 2024. 5, 6

[22] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer, 2024. 1, 2

[23] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2

[24] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. 3

[25] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2025. 5, 6

[26] Ruixiang Jiang and Changwen Chen. Diffartist: Towards aesthetic-aligned diffusion model control for training-free text-driven stylization, 2024. 6

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 1

[28] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion, 2024. 3

[29] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 6

[30] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. 3

[31] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. 7

[32] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 3, 7

[33] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. 5

[34] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 6, 7

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1, 3

[36] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 3

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6

[38] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. 6

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[40] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024. 6

[41] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. *arXiv preprint arXiv:2403.09622*, 2024. 1

[42] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 3

[43] Pengqi Lu. Qwen2vl-flux: Unifying image and text guidance for controllable image generation, 2024. 2, 3

[44] Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models, 2024. 3

[45] Jian Ma, Chen Chen, Qingsong Xie, and Haonan Lu. Pea-diffusion: Parameter-efficient adapter with knowledge distillation in non-english text-to-image generation. *arXiv preprint arXiv:2311.17086*, 2023. 1, 6

[46] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *arXiv preprint arXiv:2303.17870*, 2023. 1

[47] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 1, 7

[48] Jian Ma, Yonglin Deng, Chen Chen, Nanyang Du, Haonan Lu, and Zhenyu Yang. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models, 2025. 1

[49] 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, 2023. 5

[50] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and

Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[51] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models, 2024. 3, 4, 7

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 3

[53] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation, 2024. 5

[54] 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, 2023. 1, 3

[55] Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, and Ran Xu. Gluegen: Plug and play multi-modal encoders for x-to-image generation, 2023. 2

[56] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild, 2023. 6

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[58] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. 1

[59] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. 1

[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1

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

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 7

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 1

[64] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. 6, 7

[65] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 3, 7

[66] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[67] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 5

[68] Yichun Shi, Peng Wang, and Weilin Huang. Seededit: Align image re-generation to image editing, 2024. 1, 2

[69] Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models, 2024. 4

[70] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. 3, 7

[71] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024. 1, 3

[72] Hugo Touvron, Louis Martin, and Kevin Stone. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

[73] Kai Wang, Fei Yang, and Joost van de Weijer. Attention distillation: self-supervised vision transformer students need more guidance, 2022. 3

[74] 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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 3

[75] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024. 3

[76] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2024. 3

[77] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023. 6

[78] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue

Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 3

[79] Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing, 2024. 4

[80] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. 6

[81] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. 6

[82] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths, 2024. 3

[83] Jianan Yang, Haobo Wang, Yanming Zhang, Ruixuan Xiao, Sai Wu, Gang Chen, and Junbo Zhao. Controllable textual inversion for personalized text-to-image generation, 2023. 7

[84] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 3, 7

[85] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 1

[86] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea, 2024. 1

[87] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. Anygpt: Unified multimodal llm with discrete sequence modeling, 2024. 3

[88] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[89] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation: Improved latent consistency distillation by semi-linear consistency function with trajectory mapping, 2024. 3

[90] Yuanzhi Zhu, Xingchao Liu, and Qiang Liu. Slimflow: Training smaller one-step diffusion models with rectified flow, 2024. 3

[91] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm, 2024. 3

X2I: Seamless Integration of Multimodal Understanding into Diffusion Transformer via Attention Distillation

Supplementary Material

6. X2I's reasoning ability and multi-turn dialogue generation ability

MLLMs themselves possess certain reasoning abilities and multi-turn dialogue capabilities. X2I also inherits some of these related abilities. Fig. 4 illustrates the reasoning abilities based on text as well as image-based reasoning abilities. It is important to note here that the features fed into X2I originate from the characteristics of MLLMs' answers. Fig. 5 demonstrates X2I's multi-turn dialogue capabilities, indicating that the generated images exhibit a certain degree of coherence and fidelity. Although these capabilities only represent a small portion of the abilities inherited from MLLMs, they provide more possibilities for improving the productivity of images.

Figure 4. X2I's reasoning ability.

Figure 5. X2I's multi-turn dialogue generation ability.

7. X2I Training with Different VLMs

Fig. 6 presents a comparison of experiments adapting X2I using three different model capacities based on InternVL. The top left shows results based on the InternVL-1B model, the top right is based on InternVL-4B, the bottom left is based on InternVL-8B, and the bottom right compares the performance of the three models on PR metrics. Overall, there are two conclusions to be drawn: firstly, as the capacity of the InternVL model increases, the peak value of the comprehensive X2I metrics also rises, which to some extent reflects the scaling-up capability of the model. The second conclusion is that as the model capacity increases, the fitting speed of X2I training also improves to some extent, likely benefiting from the design of AlignNet and the support of attention distillation.

8. Consistency Analysis of Feature Distribution for Multiple Modes of MLLMs

As shown in Fig. 7, in the unaligned state, MLLMs generate heterogeneous feature distributions for text, image, audio, and video across four modalities, embedding these distributions into a unified cross-modal shared semantic space. At this stage, there is a significant domain gap between the feature distribution of the text modality and the output of T5. By performing parameter space projection on the textual output layer of MLLMs through pure text training, the textual feature distribution is partially aligned with the T5 encoder. The aligned X2I architecture inherits the multimodal association capabilities of MLLMs via the shared semantic space.

9. Acceleration of Training

Training-Inference Separation: Placing both the teacher model and the student model on the same GPU can lead to insufficient single-GPU memory, necessitating the use of ZeRO3[59]. However, ZeRO3 significantly reduces the inference speed of the teacher model and also decreases the training efficiency across multiple machines. By separating training and inference, the model that requires training and the model that performs only inference are placed on different GPUs. This setup results in very high inference efficiency for the teacher model without the need for multi-GPU communication. Meanwhile, the student model only requires data parallelism, with communication overhead being much lower than ZeRO3, greatly enhancing multi-machine training efficiency compared to ZeRO3.

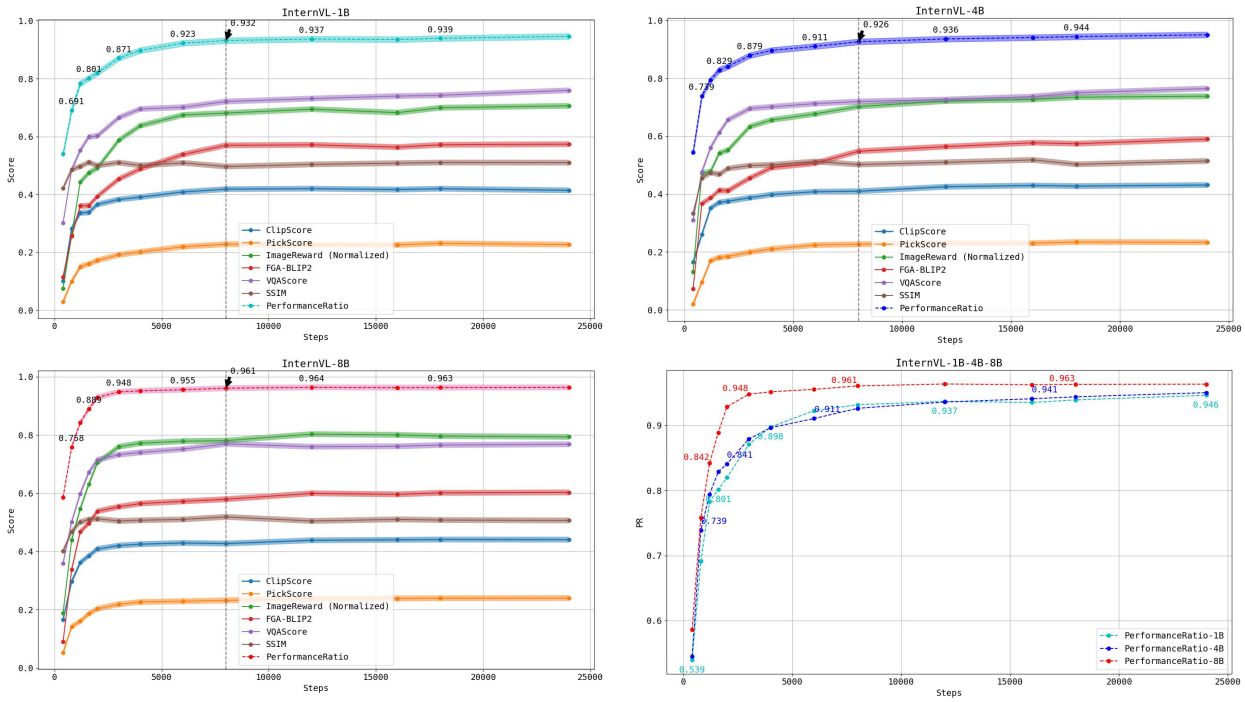


Figure 6. The Impact of varying capacities of VLMs on X2I training.

Training-Inference Overlap: By overlapping the inference of the teacher model with the training of the student model, the inference time of the teacher model can be perfectly hidden. This approach significantly improves the efficiency of distillation training.

10. Objective Performance of GenAI Bench and EvalMuse-40K

Tab. 4 displays the results from the GenAI-Bench and EvalMuse evaluation sets, with performance scores also exceeding 99%. This further demonstrates that X2I's results in T2I generation are nearly indistinguishable from those of the teacher model Flux.1.

11. Multilingual T2I Generation

Tab. 5 includes three multilingual evaluation datasets from Multilingual-General, as well as three additional translations in German, Portuguese, Spanish, and French. While the performance in English remains consistent with a one percentage point difference similar to before, the performance in other languages varies with some degree of decline. There are two main reasons for this difference:

The training corpus for QwenVL is primarily sourced from Chinese and English, with relatively fewer resources for other languages, making alignment with English datasets challenging. Errors introduced by the translation software, as all eight evaluation metrics are tested on English evaluation datasets, necessitating a translation back to English for non-English languages. In comparing with existing open-source models in the industry, such as PEA-Diffusion and Sana, the multilingual T2I models in this study demonstrate significant advantages. PEA-Diffusion is based on the SDXL model, while our comparison involves the official open-source "MultilingualFLUX.1". Sana is compared with the official open-source "Sana_1600M_1024px_MultiLing", with most of the Chinese T2I generation results failing, hence the lack of presentation of objective results for Chinese.

12. IT2I and I2I Comparison of Objective Indicators

We evaluate the ClipScore, PickScore, and ImageReward. The input prompts consist of the reference image prompt combined with editing prompts. For an objective test, we focus on common editing tasks such as addition, specifically adding five common animals (cat, dog, lion, tiger, elephant) to the original image. This evaluation primarily assesses the image instruction editing ability. For I2I metrics, we also provide SSIM scores. The corresponding prompt is the prompt of the reference image itself. The reference images are generated by the teacher model corresponding to

Evaluation	Model	ClipScore	PickScore	HPSv2	ImageReward	FGA-BLIP2	GPT-4o	VQAScore	BLIP-VQA	Performance
GenAI-Bench	Flux.1	0.4633	0.2907	0.2355	1.3105	3.6218	0.8195	0.7817	0.4970	99.13%
	X2I	0.4571	0.2873	0.2458	1.2536	3.5736	0.7925	0.7657	0.4957	
EvalMuse-40K	Flux.1	0.4554	0.2575	0.2128	1.1721	3.4441	0.7980	0.7168	0.3627	99.09%
	X2I	0.4424	0.2482	0.2134	1.1315	3.4648	0.7718	0.7038	0.3559	

Table 4. Objective Performance on the GenAI-Bench and EvalMuse Evaluation Dataset.

Language	Model	ClipScore	PickScore	HPSv2	ImageReward	FGA-BLIP2	GPT-4o	VQAScore	BLIP-VQA	Performance
English	Flux.1	0.4517	0.2578	0.2433	1.3891	3.7789	0.9105	0.8741	0.4101	-
	PEA-Diffusion	0.4325	0.2485	0.2401	1.1584	3.5987	0.8712	0.8501	0.3602	96.90%
	X2I	0.4543	0.2577	0.2445	1.4043	3.7568	0.8905	0.8667	0.3905	**99.25%**
Chinese	PEA-Diffusion	0.4252	0.2415	0.2412	1.0582	3.5542	0.8542	0.8469	0.3312	95.67%
	X2I	0.4481	0.2552	0.2429	1.2998	3.6458	0.8614	0.8601	0.3471	**97.65%**
German	PEA-Diffusion	0.4101	0.2202	0.2417	0.9854	3.2546	0.8498	0.8201	0.3042	93.33%
	Sana	0.4154	0.1929	0.2315	0.4168	2.9490	0.8045	0.7954	0.2855	-
	X2I	0.4356	0.2368	0.2419	1.1736	3.5350	0.8715	0.8412	0.3178	**96.03%**
Portugal	PEA-Diffusion	0.3952	0.2196	0.2399	1.0123	3.4550	0.7854	0.7952	0.3020	92.68%
	Sana	0.4253	0.1958	0.2217	0.5240	3.4910	0.7905	0.8014	0.3125	-
	X2I	0.4344	0.2385	0.2427	1.1610	3.5892	0.8221	0.8254	0.3345	**95.57%**
Franch	PEA-Diffusion	0.4054	0.2285	0.2347	0.8975	3.2541	0.8014	0.8077	0.2998	92.20%
	Sana	0.4346	0.2080	0.2301	0.7202	3.1522	0.7899	0.8158	0.3101	-
	X2I	0.4322	0.2353	0.2415	1.0956	3.4681	0.8317	0.8298	0.3210	**94.91%**
Spain	PEA-Diffusion	0.4051	0.2142	0.2298	1.0214	3.2121	0.8012	0.7946	0.3014	92.06%
	Sana	0.4199	0.1977	0.2145	0.5710	3.0526	0.7854	0.7714	0.2845	-
	X2I	0.4320	0.2363	0.2427	1.1737	3.5398	0.8223	0.8399	0.3311	**95.36%**
Russia	PEA-Diffusion	0.4012	0.2154	0.2314	0.9214	3.2985	0.7988	0.7901	0.2952	91.84%
	Sana	0.3641	0.1479	0.2157	-0.2139	2.5536	0.7012	0.7265	0.2545	-
	X2I	0.4312	0.2354	0.2428	1.1475	3.5254	0.8237	0.8412	0.3236	**94.94%**
Itali	PEA-Diffusion	0.3925	0.2101	0.2351	0.9215	3.2541	0.7952	0.7892	0.2912	91.47%
	Sana	0.4124	0.1886	0.2201	0.4082	2.9347	0.7524	0.7412	0.2641	-
	X2I	0.4286	0.2327	0.2419	1.0956	3.4769	0.8285	0.8214	0.3183	**94.69%**

Table 5. Multilingual T2I Performance on the Multilingual-General Evaluation Dataset.

200 English corpus from the Multilingual-General dataset. This evaluation mainly examines the model's capability to generate variations of images. The baseline model for comparison is Qwen2VL-Flux.

IT2I and I2I. Tab. 6 presents metrics for IT2I and I2I, comparing the X2I model with Qwen2VL-Flux. In all results, the metrics of X2I are significantly higher than those of Qwen2VL-Flux, indicating that X2I have a strong semantic understanding of images and performs well in common image generation tasks involving instruction editing.

13. Plug-and-Play

Benefiting from the alignment strategy of attention distillation, the MLLM encoder acquired by X2I can be directly transferred to Flux.1 variant models and downstream ecosystems without requiring retraining.

13.1. Transfer Capability on Flux.1 Variant Models

By replacing the text encoders of the sampling acceleration model Flux.1-Turbo, aesthetic fine-tuning model Flux.1-Shuttle3.1, and compression model FLEX with X2I's MLLM as the encoder, we derived corresponding variants X2I-Turbo, X2I-Shuttle, and X2I-FLEX. These variants are systematically compared with the original models in text-to-image generation performance. As demonstrated in Fig. 33, Fig. 34, and Fig. 35, our alignment strategy achieves robust cross-architecture transferability.

13.2. Downstream Task Transferability

The aligned MLLM encoder can be directly transferred to downstream ecosystems including LoRA, inpainting, outpainting, ControlNet, and IP-Adapter without architectural modifications. As demonstrated in Fig. 36, Fig. 37, and Fig. 38, X2I achieve comparable or superior performance on children's simple sketch LoRA, outpainting, and inpainting tasks.

ControlNet. For ControlNet-related tasks, X2I's utilization of the MLLM as an encoder enables original image input support, thereby capturing finer-grained visual details. As shown in Fig. 40, X2I achieves enhanced performance under canny, depth, and hed reference conditions. We also compared the performance of ControlNet downstream models based on Flux.1 directly applied to the X2I framework as Tab. 7 shows. X2I w/i indicates that the reference image is also input to the X2I framework during inference. The results show that including the image can further enhance

Model	I2I				TI2I		
	ClipScore	PickScore	ImageReward	SSIM	ClipScore	PickScore	ImageReward
Qwen2VL-Flux	0.3919	0.1845	0.6824	0.4045	0.3772	0.1429	-0.3914
X2I	**0.4347**	**0.2371**	**1.2446**	**0.5208**	**0.4318**	**0.182**	**0.732**

Table 6. Performance of X2I in IT2I and I2I Generation.

the performance on relevant tasks based on three metrics.

ControlNet	Canny Edge (F1 Score↑)	Hed Edge (SSIM↑)	Depth Map (RMSE↓)
Original	0.2903	0.5025	53.79
X2I	0.3144	0.5340	49.09
X2I w/i	**0.3304**	**0.5377**	**43.03**

Table 7. Performance of X2I Framework in Transfer to ControlNet Downstream Tasks.

IP-Adapter. In IP-Adapter tasks in Fig. 39, X2I exhibits improved fidelity preservation owing to its original image encoding capability. Similarly, we compared the performance of the IP-Adapter downstream model based on Flux.1 directly applied to the X2I framework as Tab. 8 shows. The evaluation dataset used is DreamBench. Without including image inputs, X2I shows a certain advantage in DINO and CLIP-I metrics, but slightly decreases. X2I w/i, with image inputs, exhibits a significant advantage in DINO and CLIP-I, but sacrifices semantic adherence capabilities as seen from CLIP-T. Similar conclusions can be drawn from the comparison table.

IP-Adapter	DINO	CLIP-I	CLIP-T
Original	0.768	0.803	**0.322**
X2I	0.779	0.811	0.320
X2I w/i	**0.863**	**0.865**	0.308

Table 8. Performance of X2I Framework in Transfer to IP-Adapter Downstream Tasks.

14. Template for X2I's input

X2I have designed an input template for the student model during the alignment training process as follows: {"text prompt":"","editing prompt":"","image prompt":"",,"video prompt":"",,"audio prompt":""},The "text prompt" represents the original prompt for generating images, the "editing prompt" is the prompt for editing instructions for the input image, video, or audio, and the "image prompt," "video prompt," and "audio prompt" respectively signify whether there is information from an input image, video, or audio. The motivation behind designing such a general template is to provide a paradigm input for subsequent stage training, LoRA training, and more controllable post-training.

15. LoRA for IT2I

Currently, the prevalent approach in the industry for generating visual images involves training models based on textual conditions as input. There is a lack of LoRA training for scenarios where the user's input involves highly personalized reference images for image-to-image and image-to-text generation. For instance, taking "style information from the reference image alone" as an image edit prompt, the training of the general X2I framework primarily captures a global semantic understanding of the image, making it challenging to individually parse and handle finer-grained layered features of the image. To address this issue, we can train a conventional LoRA module within the X2I framework to solve these kinds of problems.

This paper trained the LoRA module for IT2I. We collected 100 abstract artistic sketches and created 100 data pairs with the input template {"text prompt": "", "editing prompt": "Style Information Only from Reference Image", "image prompt": "yes", "video prompt": "no", "audio prompt": "no"}. Here, the text prompt is set as the corresponding textual description for each of the 100 images collected, and the "image prompt" is randomly selected from the 100 data pairs.

Fig. 8 shows the comparison of results before and after LoRA training. The conclusion indicates that LoRA training can prompt the X2I framework to selectively learn and extract different semantic information from input images.

16. X2I enhanced Training Results

We further present a comparison of Van Gogh style Fig. 9 and Cubism style Fig. 10. Overall, the semantic consistency ability of X2I shows some improvement compared to DiffArtist, and the fidelity to the reference image significantly improves compared to X2I trained without enhancement.

17. X2I Generation

17.1. Text-to-Image

Fig. 12 compares the effects of Flux.1 and X2I in generating images for complex prompts , and also Fig. 13 demonstrates X21's ability to generate images in multiple languages, including German, English, French, Japanese, Thai, Vietnamese, and Chinese.

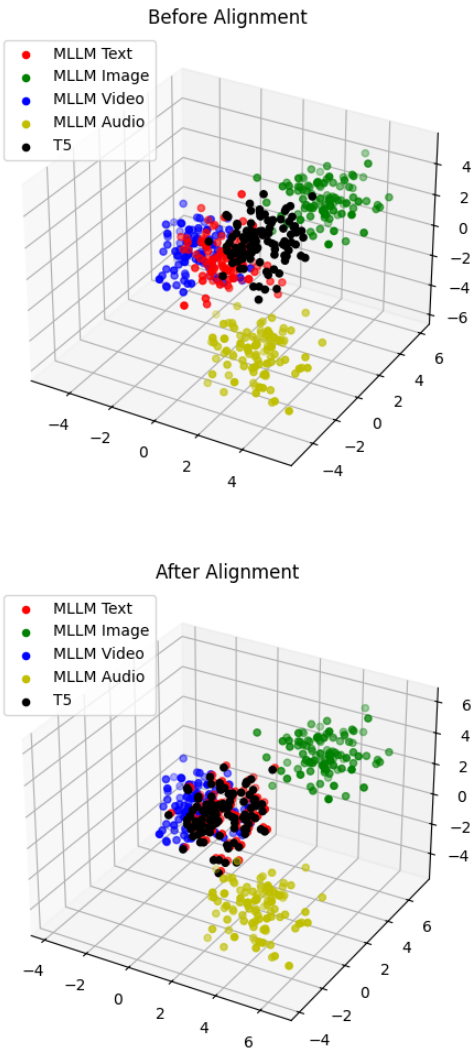


Figure 7. Schematic diagram of alignment process for different modalities.

17.2. Image-to-Image

MLLM empowers X2I with the capability to understand both single and multiple images, enabling it to perform reference-guided image generation in Fig. 14, celebrity and IP generation in Fig. 15, and multi-image composition tasks in Fig. 16, Fig. 17, Fig. 18 and Fig. 19.

17.3. Text+Image-to-Image

X2I demonstrates capabilities including user-prompt-driven style transfer, expression editing as shown in Fig. 20, Fig. 21 and Fig. 23, along with single image or multi-image editing and fusion tasks illustrated in Fig. 22, Fig. 24, and Fig. 25. Furthermore, leveraging MLLM's robust OCR capacity, the system generates images through direct interpretation of visual content in input images while supporting multilingual visual generation, evidenced by the results in Fig. 26.

17.4. Video-to-Image

MLLM possesses video comprehension capabilities that enable X2I to directly generate images based on the semantic content of input video sequences, as shown in Fig. 27. X2I can still generate high-resolution images through the prompt of the low-resolution video.

17.5. Audio-to-Image

Leveraging the audio comprehension capabilities of MLLMs such as MiniCPM-o, after alignment, X2I can directly generate images based on music with lyrics, instrumental music, and natural sounds as shown in Fig. 28, Fig. 29 and Fig. 30. All audio is fed into X2I without any prior processing.

17.6. X-to-Image

As demonstrated in Fig. 31, X2I can comprehend hybrid inputs combining audio, images, videos, and text prompts to generate images. The rendering of the first row has its input text description as "In the abandoned cyberpunk ruins, a mysterious symbiotic relationship has been established between nanobots and humans. These nanobots zip back and forth through the dilapidated buildings, repairing the decaying parts of the city, while humans, by merging with the nanobots, gain extraordinary abilities. In this sci-fi world, a dazzling yet perilous new ecosystem is created, captivating people in the fantasies of the future.". Moreover, as shown in Fig. 32, when the same video is used as a prompt, accompanying it with music produces distinct effects, demonstrating X2I's comprehension of multimodal prompts.

18. Limitations

X2I demonstrates some limitations in precise and controllable image instruction editing in a unified framework. For instance, the success rate of removing a specific element from an image is relatively low. The MLLM component of the X2I framework struggles to parse the input image into fine-grained semantic layers according to textual instructions and then output the corresponding intermediate-level features. Additionally, the LightControl have only been validated in the domain of image stylization, and further

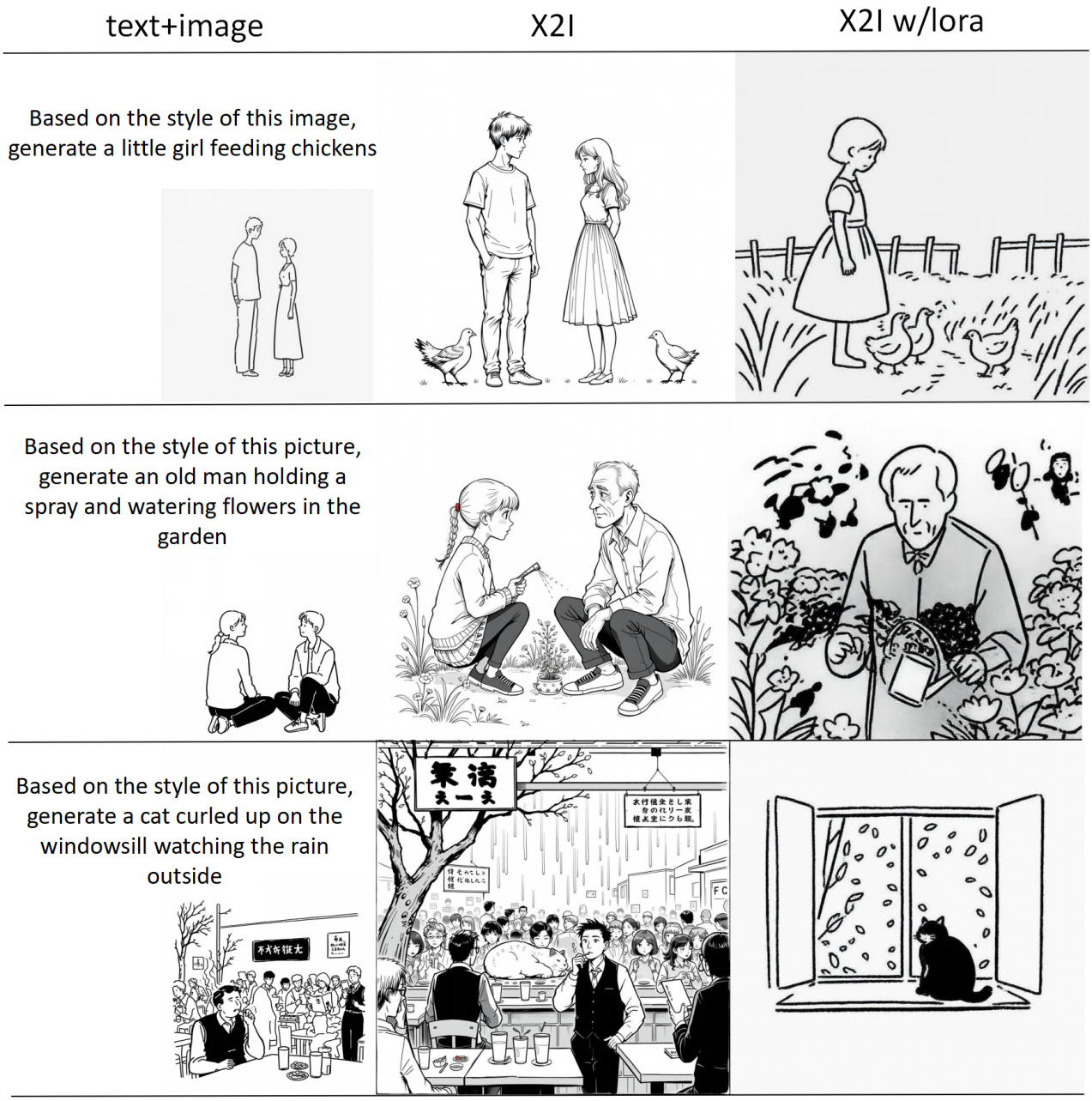


Figure 8. The effects before and after LoRA training. The conclusion indicates that LoRA training can encourage the X2I framework to selectively learn and extract different semantic information from input images.

exploration is needed for more extensive image instruction editing such as addition, deletion, and modification. Moreover, X2I have not fully exploit the multimodal understanding capabilities of MLLMs, such as summarizing long texts, logical reasoning, few-shot learning, and CoT capabilities. Future directions should focus on more intelligent and precisely controllable visual generation. Building high-quality datasets for various image instruction editing tasks is crucial, as well as leveraging these datasets to enhance the comprehensive capabilities of X2I.

19. Acknowledgements

We extend our heartfelt gratitude to the dedicated researchers and scholars whose pioneering work in the fields

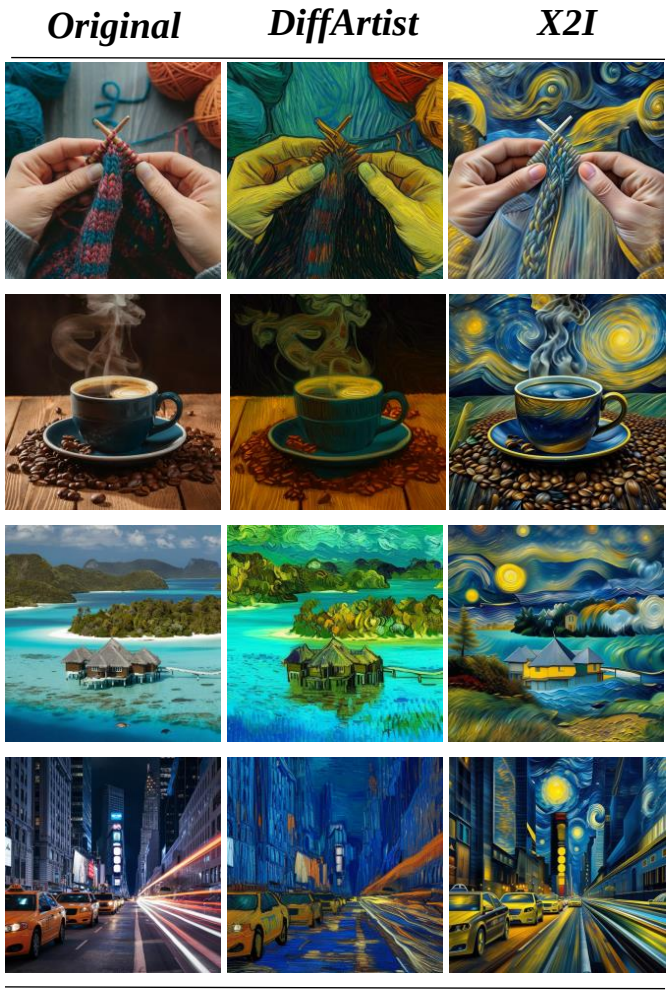


Figure 9. Comparison between DifffArtist and X2I Van Gogh style transfer capability.

of MLLMs and T2I generation has significantly contributed to the foundation and advancement of this study. In particular, we want to acknowledge the open-source community behind Flux.1, whose groundbreaking models have greatly facilitated our research. Your collective expertise and commitment to open science have been a constant source of inspiration and an indispensable resource for our work. Thank you for your unwavering dedication to pushing the boundaries of T2I and for fostering an environment of collaboration and innovation.

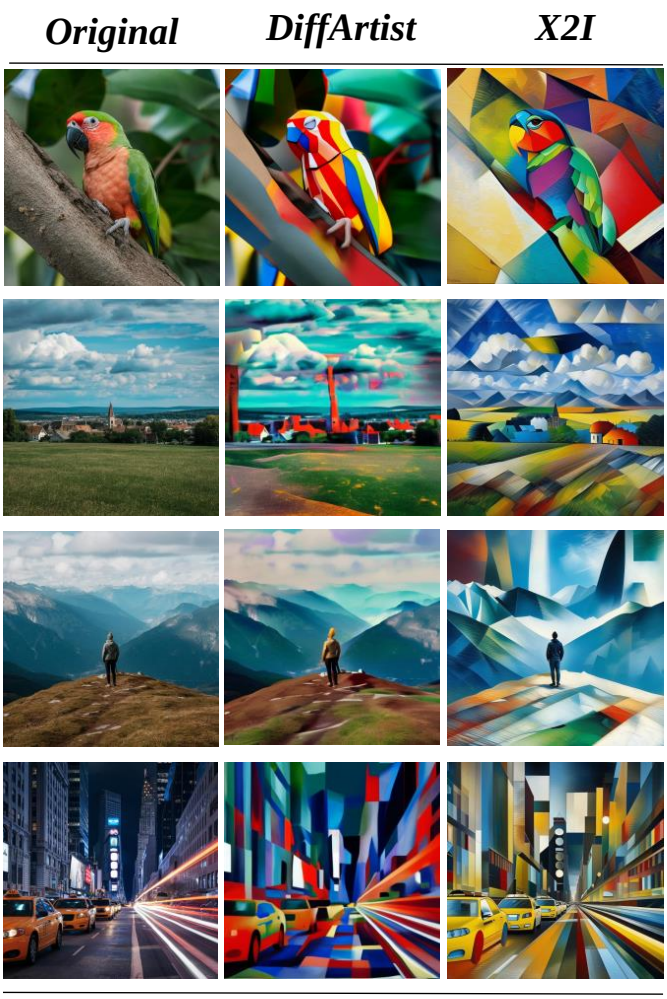


Figure 10. Comparison between DifffArtist and X2I cubism style transfer capability.

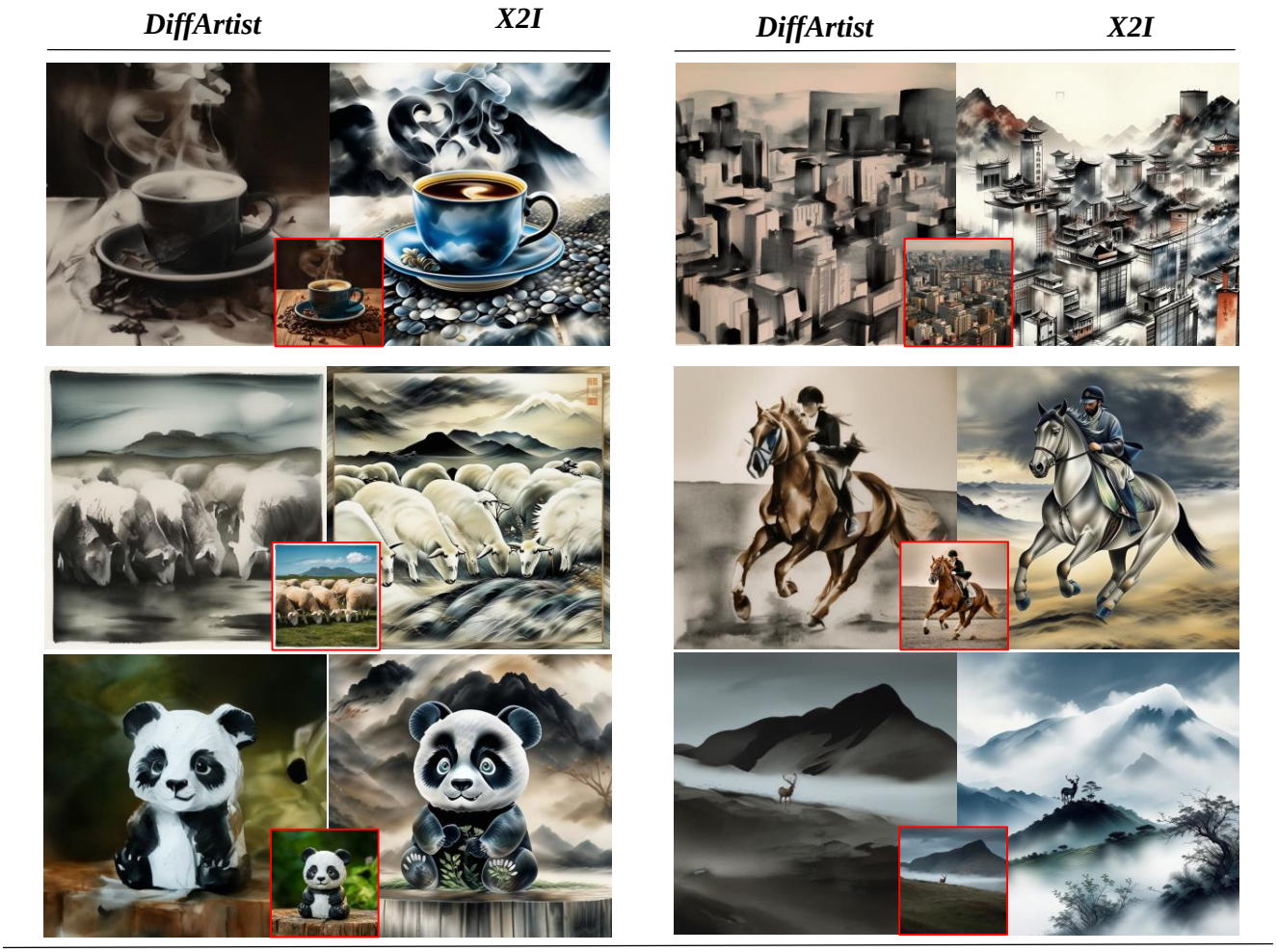


Figure 11. Comparison between DifffArtist and X2I ink painting style transfer capability.

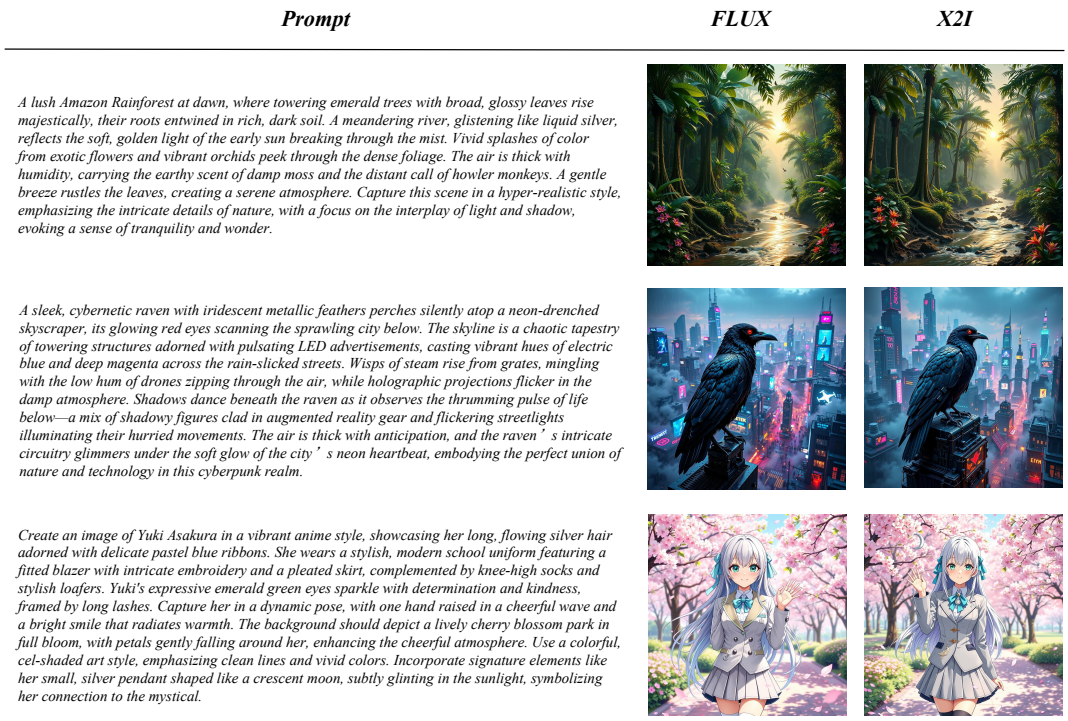


Figure 12. Comparison of the image generation results using the same prompt between the original FLUX.1 and our X2I.

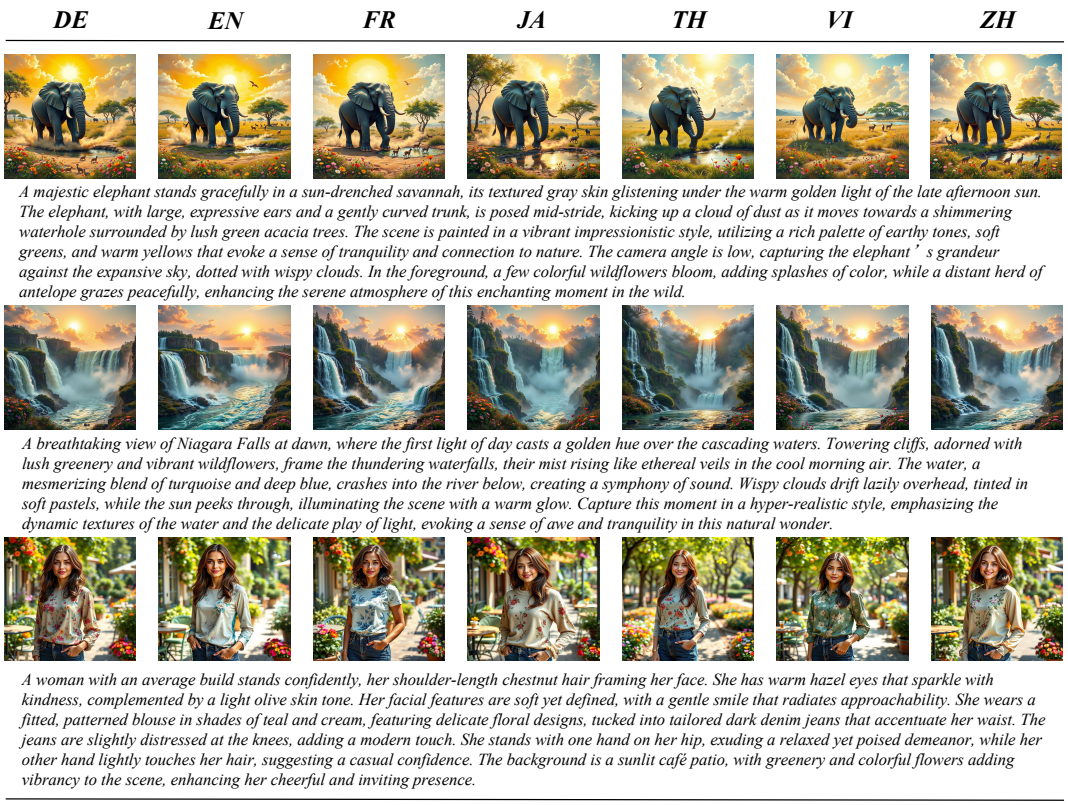


Figure 13. Comparison of text-to-image generation results using different languages.

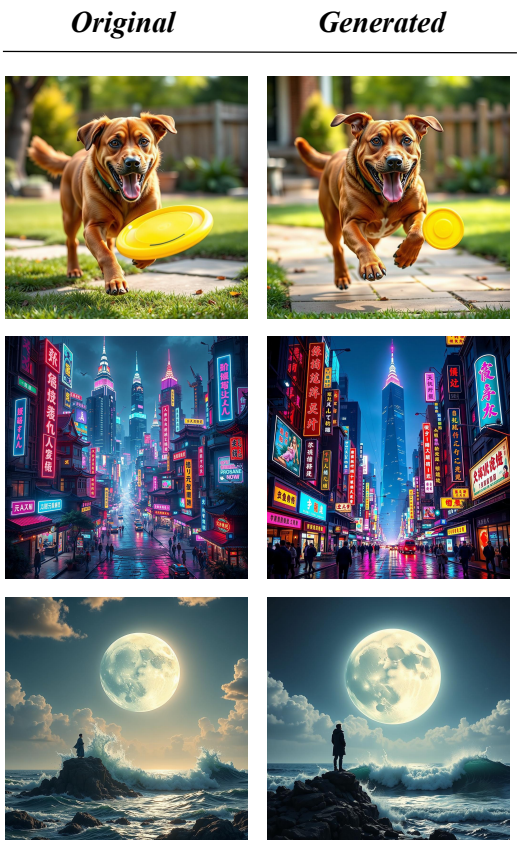


Figure 14. Image-to-image generation.

Figure 15. Celebrity or IP image-to-image generation.

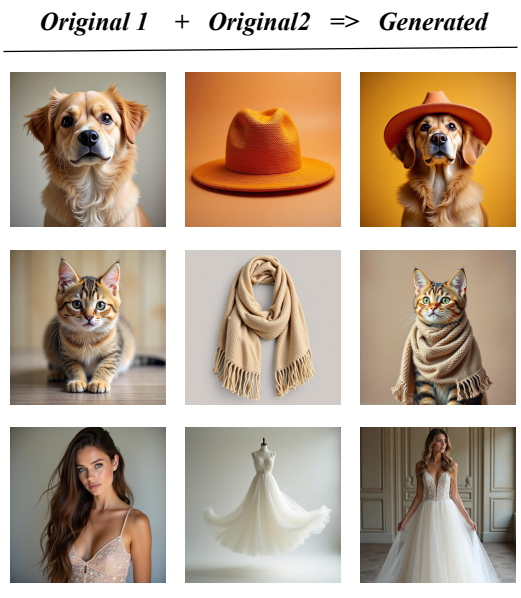


Figure 16. Clothing and accessories multi-image generation.

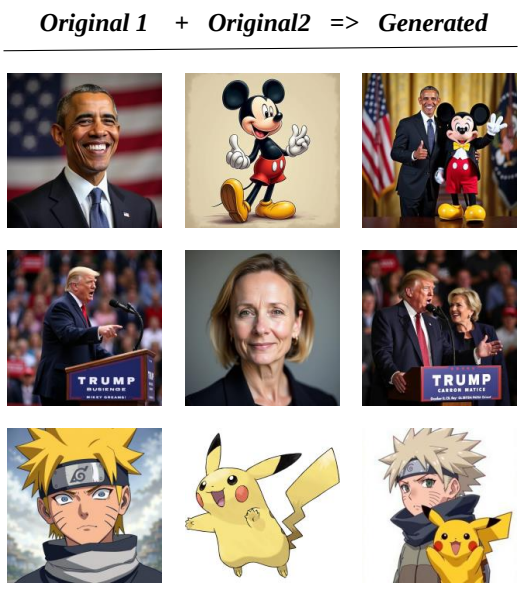


Figure 17. Group photo multi-image generation.

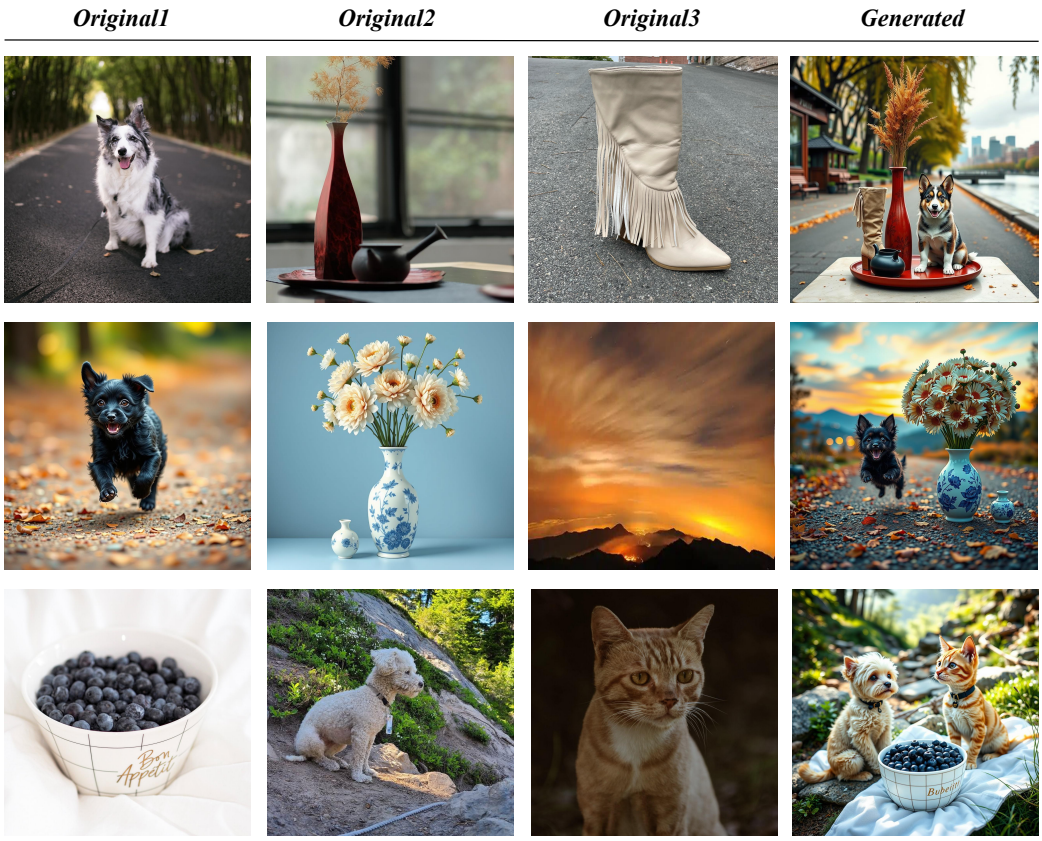


Figure 18. Blend three images to generate a new image.

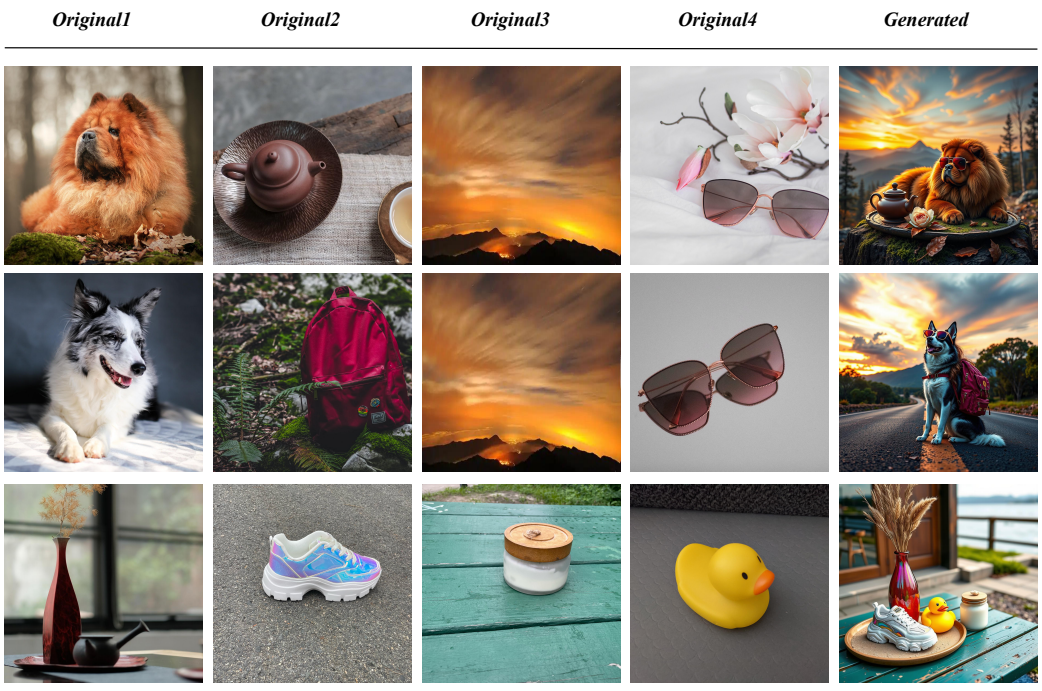


Figure 19. Blend four images to generate a new image.

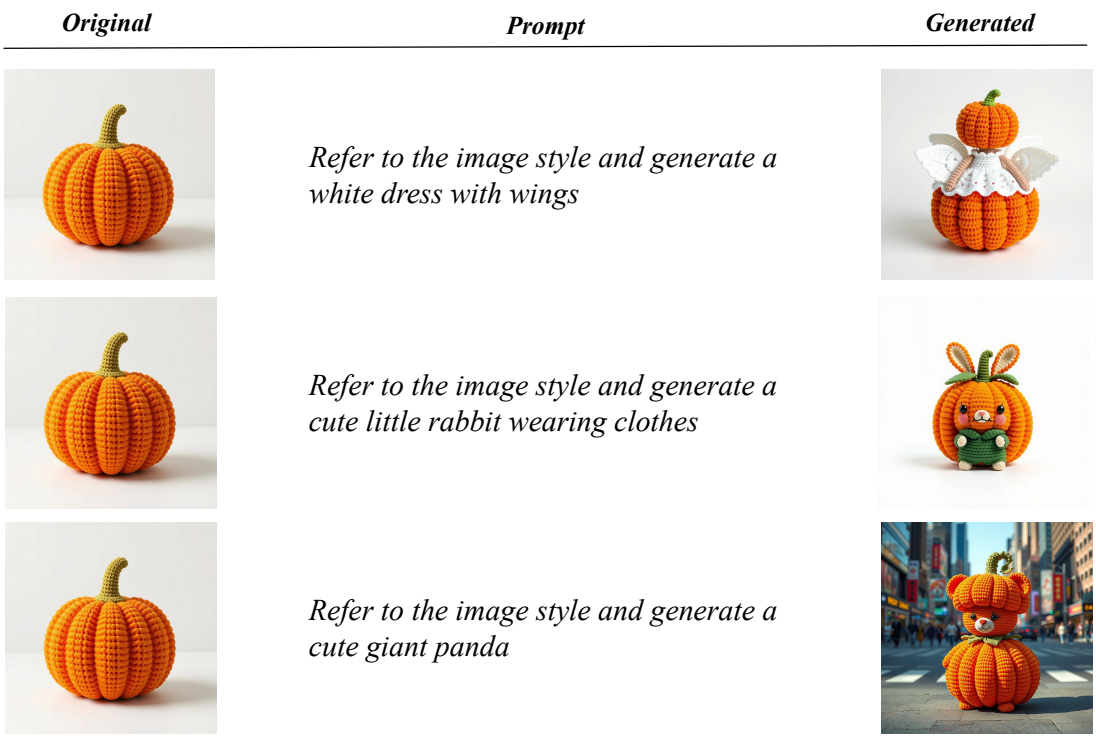


Figure 20. Image style transfer, refer to the image style and generate an image based on the prompt.

Figure 21. Style Editing, change the style of the specified image.

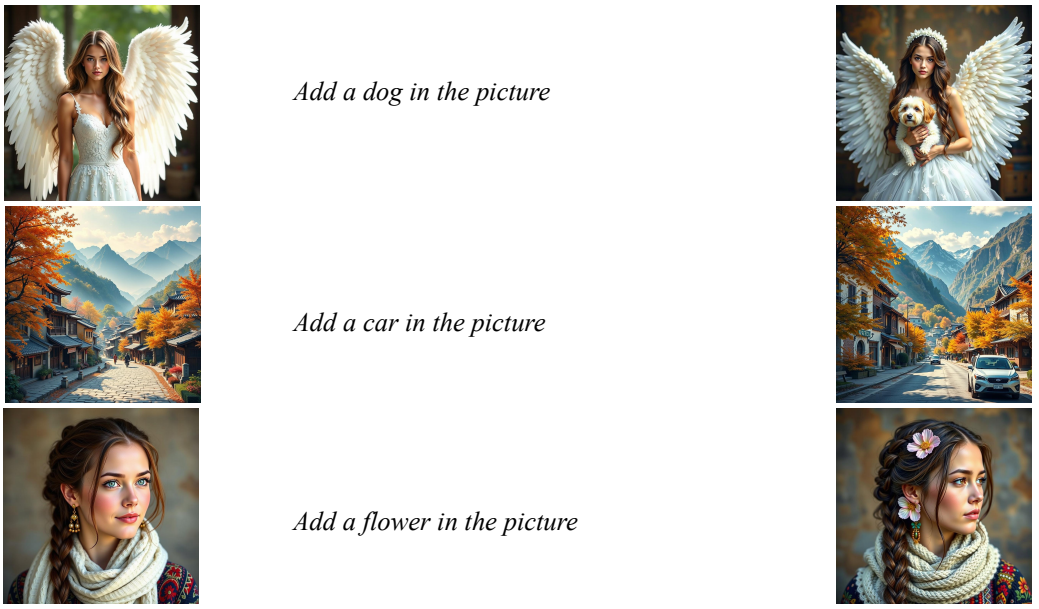


Figure 22. Image editing, add a new object to the image.

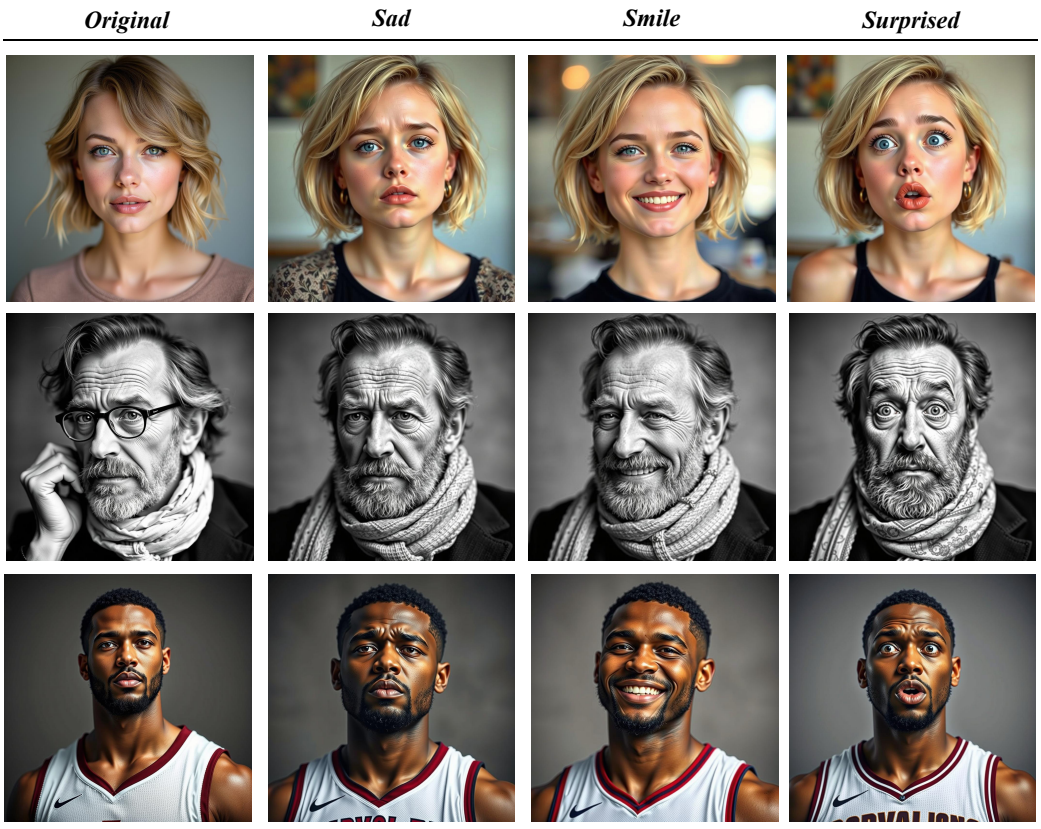


Figure 23. Expression editing, change the expression of a person in the image.

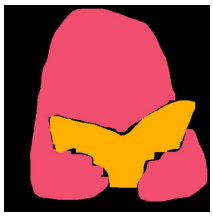
Original	Prompt	Generated
	*A berry_bowl with a blue house in the background.*	
	*Place the cup from the image onto a table with a potted plant.*	
	*A lion wearing the clothes from the image is reading a book.*	

Figure 24. Single image and prompt fusion for image generation.

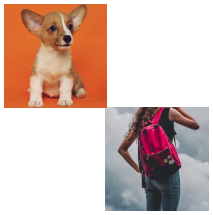
Original Images	Prompt	Generated
	*With a tree and autumn leaves in the background*	
	*With the Eiffel Tower in the background*	
	*The <image3> is wearing a <image1> and riding a <image2> on the <image4>, the <image2> is also wearing a <image1>*	

Figure 25. Multiple image and prompt fusion for image generation, in the third case, in order to express the complex relationship between multiple entities, we generate images by embedding image tokens between text.

Original	*Prompt*	*Generated*
奇幻的金属树在无尽的半透明地平线上拔地而起，树干散发着幽蓝的光芒，仿佛每个分支都在进行一场静谧的电子交响乐。天空中漂浮着旋转的三角形云朵，这些云朵都是由光子编织而成，不断地散发炫目的细光，如同无数数量子跃迁的瞬间闪烁。远处，星际菱形飞船在无形的轨道上翩然舞动，尾部拖拽着长长的等离子流，色彩从鲜艳的紫罗兰渐变至深邃的翡翠绿。 地表覆盖着银色的纳米群落，每个粒子都在执行着对周围环境的微调整和修复任务，这些粒子时而汇聚成镜面般的河流，时而散落成微小的反射光点，如同白昼中的星空。竟有人影穿梭于其中，那是数个由超导纤维编织而成的仿生体，他们的动作如行云流水，却带着某种机械的精准。他们手中的反重力工具如同古老的法杖，但每一次挥动都能在空中画出绚丽的尾迹，仿佛操纵着看不见的能量网络。 而在整个场景的正中央，一栋如梦似幻的建筑伫立着，表面布满着动态的全息图案，闪烁的符号和图形不断变幻，仿佛在传递着某种古老的智慧与现代的哲理。城市的呼吸，通过那些如诗般的电磁韵律，传递至整个光波交织的苍穹。	*OCR text recognition*	
Chapter 2: Imagine a floating castle made entirely of translucent crystal, suspended above an endless ocean of molten gold. The sky is a tapestry of swirling, iridescent clouds that glimmer with hues of ultraviolet and deep indigo. Gigantic, bioluminescent jellyfish drift peacefully through the air, their long tendrils entwined with clusters of luminescent fireflies. On a distant horizon, jagged mountains carved from sapphire and emerald rise majestically, their peaks piercing through the fabric of the sky itself and revealing glimpses of countless galaxies beyond. A lone, ancient tree with branches made of shimmering glass grows on a floating island, its leaves like stained glass windows reflecting the myriad colors of the surreal surroundings. Standing on a bridge of golden light that connects the tree to the crystal castle, a figure cloaked in living shadows gazes out upon the scene, holding a staff that emanates a soft, pulsating glow. The water beneath is alive with ethereal, spectral fish that leave trails of sparkling stardust in their wake, and the air is filled with a symphony of otherworldly whispers and haunting melodies.	*OCR text recognition*	
Imagine un vasto bosque de flores collage, où le sol is made of floating puzzles pieces that are colorful y con trozos that nunca fit together. Among the arbres gigantesques, des luciérnagas éclatantes fliegen en zigzag zwischen las hojas doradas. Fliegende Fische mit brillanten Schuppen passent doucement أسماك الألوان forming whimsical patterns in the air. In the middle of the scene, un grand reloj avec hands that point in all directions y cada segundo se vuelve más kunstlerisch. On top of this puzzling landscape, un zorro avec des ailes transparentes y alas de mariposa está tocando un violín des réflexions dorées. Todo su entorno se mueve too, como si the entire world respirara juntos. Kleine Elfen mit leuchtenden Flügeln scattered around him, dancing al ritmo de una música invisible et enchantée. Furthermore, caracoles gigantes llevan maisons de lumière colorée, écrivant des poèmes mystérieux με φωτεινά ίχνη πίσω τους. Mientras tanto, the ciel noir is scattered con des estrellas brillantes que parpadean lentamente. Un énorme chat noir avec des yeux lumineux βρίσκεται en el horizonte, son regard fixant un punto no claro con une intensité énigmatique. En este sueñoscape, every object and creature coexists in a harmonic chaos, relays silences mysterious that fuse seamlessly, transforming the scene into a surrealist symphony of wonders. Quisque elementum in this universus alternate is an embodiment of imagination gone wild, transcending the boundaries of logical harmony while celebrant l'essence of creative liberty.	*OCR text recognition*	
A young woman wearing a green coat and a yellow hat, holding a book called 'Prophet'. The cover is blue, and the silhouette in the book is a silhouette of a male head. The background is red.	*OCR text recognition*	

Figure 26. Image generation from screenshot inputs with OCR recognition.

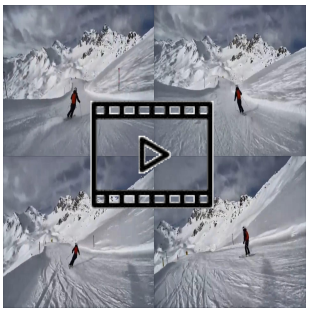
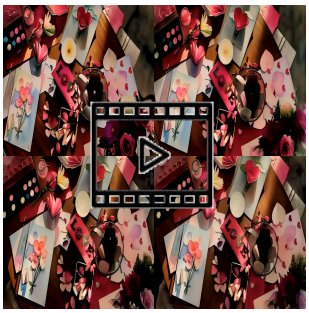
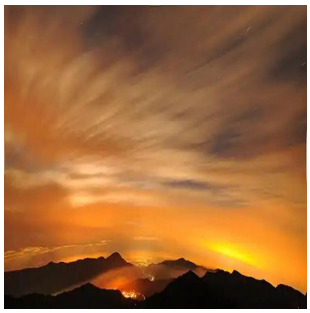
Original Video	*Original Image*	*Prompt*	*Generated*
	*None*	*None*	
	*None*	*None*	
	*None*	*High resolution*	
		*Wearing scarf and hat*	

Figure 27. Video-to-image generation, generate images based on the content of the video.

Original Song Audio	***Generated***
"Hotel California" by Eagles *Lyrics: "On a dark desert highway, cool wind in my hair, warm smell of colitas rising up through the air..."*	
"What a Wonderful World" by Louis Armstrong *Lyrics: "I see trees of green, red roses too, I see them bloom for me and you, and I think to myself, what a wonderful world..."*	
"Over the Rainbow" by Judy Garland *Lyrics: "Somewhere over the rainbow, way up high, there's a land that I heard of once in a lullaby..."*	

Figure 28. Music-to-image generation, generate images from music with lyrics.

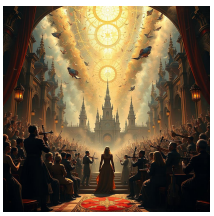
Original Classical Music Audio	***Generated***
"Moonlight Sonata" by Ludwig van Beethoven	
"Rondo alla Turca" by Wolfgang Amadeus Mozart	
"A Dream of Wedding" by Richard Clayderman	

Figure 29. Instrumental-music-to-image generation, generate images from instrumental music audio.

Original Audio	***Generated***
The sounds of insects and birds	
Train whistle	
Sound of flowing water	
And speaking of the Irish way, have you ever tried explaining hurling to someone who's never seen it? It's like trying to describe the rules of Quidditch, but with less flying and more shouting. Ah, but sure, it's all in good fun.	

Figure 30. Audio-to-image generation, generate images from natural sounds, vehicle sounds, and complex human sounds.

Original Image/Video + Music Audio	***Prompt***	***Generated***
 *Mechanical operation*	*Science fiction novel paragraph*	
 *Electronic music with strong rhythm*	*None*	
 *Moonlight Sonata for Piano*	*Artistic conception poetry*	

Figure 31. X-to-image generation.

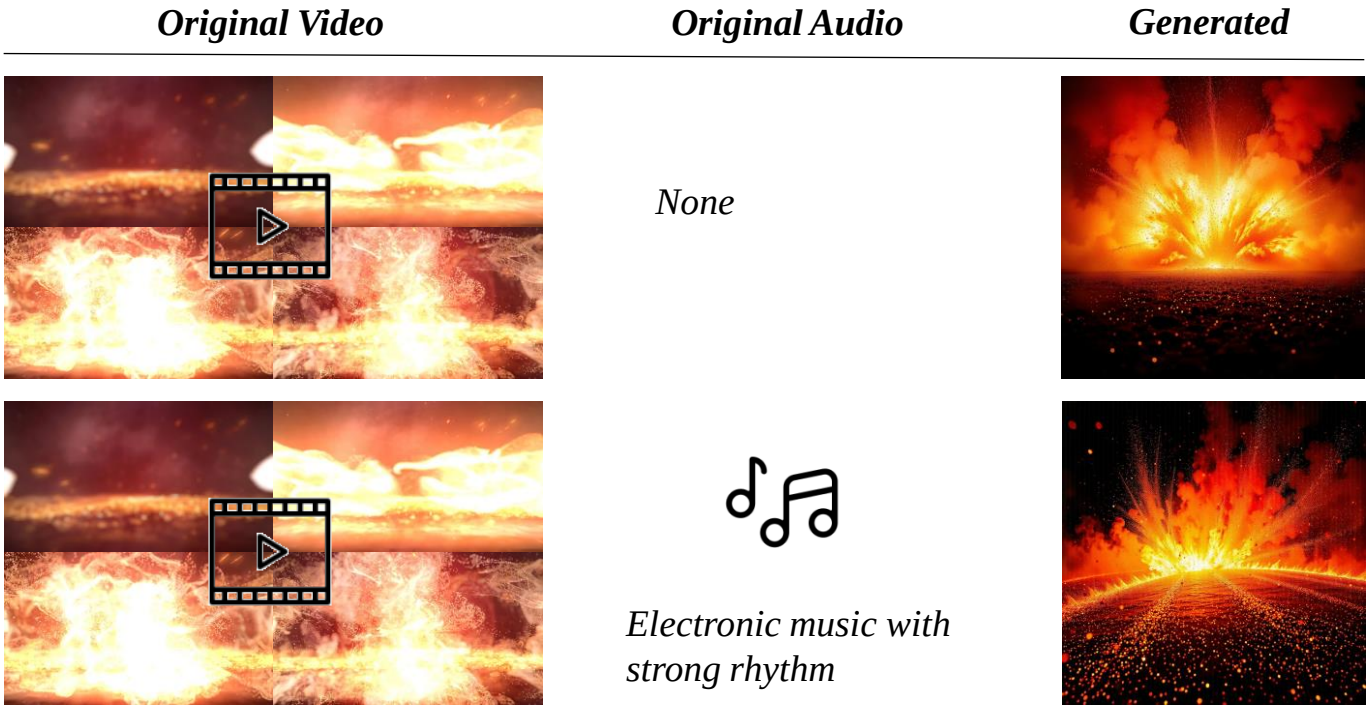


Figure 32. Images generated by combining the particle collision video with electronic music exhibit a greater sense of particle texture and oscillation compared to those created using just the video alone.

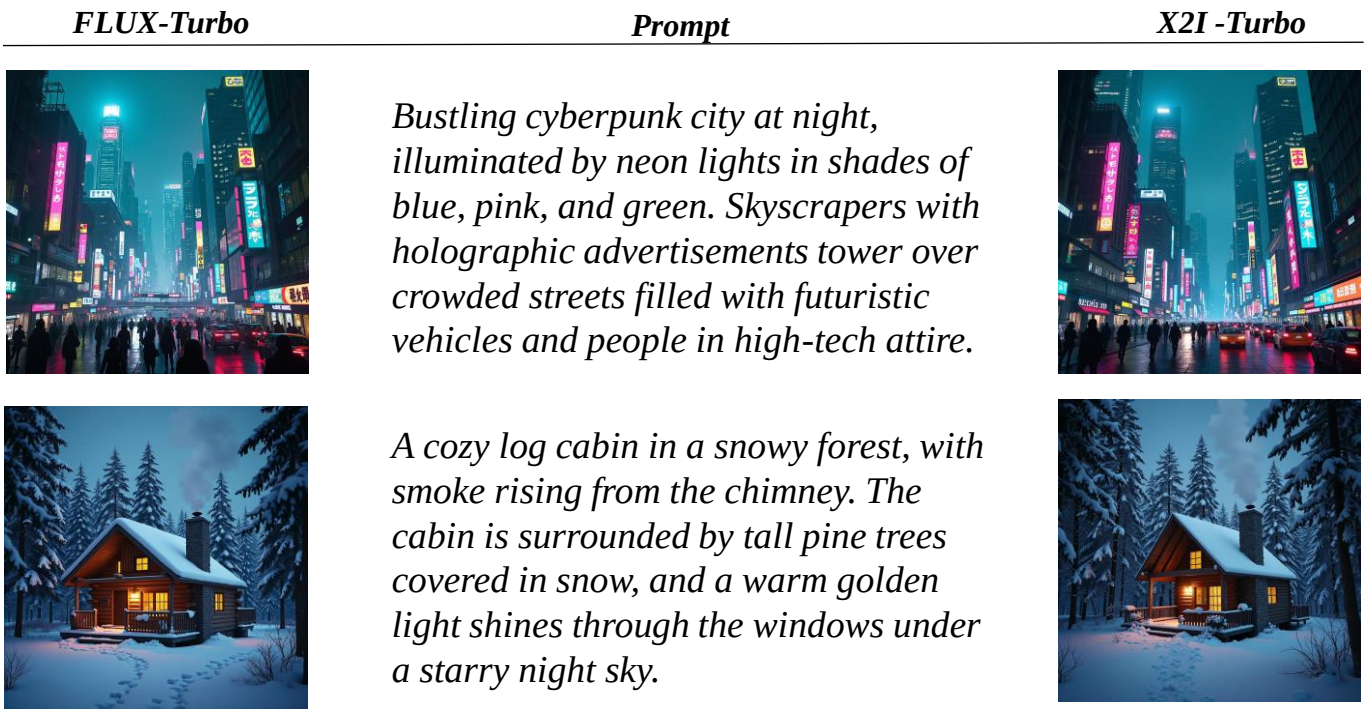


Figure 33. Comparison of the sampling acceleration model FLUX-Turbo and the X2I-Turbo obtained by replacing the text encoder of FLUX-Turbo with MLLM.

FLUX-Shuttle	***Prompt***	***X2I -Shuttle***
	*A peaceful beach at sunset with golden sand, calm waves, and a colorful sky.*	
	*A bustling city park with green trees, walking paths, and people enjoying the day.*	

Figure 34. Comparison of the aesthetic fine-tuning model FLUX-Shuttle3.1 and the X2I-Shuttle obtained by replacing the text encoder of FLUX-Shuttle with MLLM.

FLEX	***Prompt***	***X2I-FLEX***
	*A highly detailed humanoid robot standing in a high-tech laboratory. The robot has a sleek metallic body with glowing blue eyes, surrounded by advanced machinery and holographic displays.*	
	*A serene fantasy landscape with a crystal-clear lake reflecting towering mountains, lush green forests, and a vibrant sunset with hues of orange and purple. A small wooden cabin sits by the shore, surrounded by blooming flowers.*	

Figure 35. Comparison of the compression model FLEX and X2I-FLEX obtained by replacing the text encoder of FLEX with MLLM.

FLUX + Lora	*Prompt*	*X2I + Lora*
	*A joyful girl with balloons floats above a city wearing a hat and striped pants*	
	*Some happy children stand ready to take pictures*	

Figure 36. Comparison between FLUX.1 with children simple sketch LoRA and X2I with children simple sketch LoRA.

Original	*FLUX + Outpainting*	*X2I + Outpainting*
		
Prompt: a picture of a candle and a book on a table.		
		
Prompt: a laptop and a light on a table.		

Figure 37. Comparison between FLUX.1 and X2I on the outpainting task.

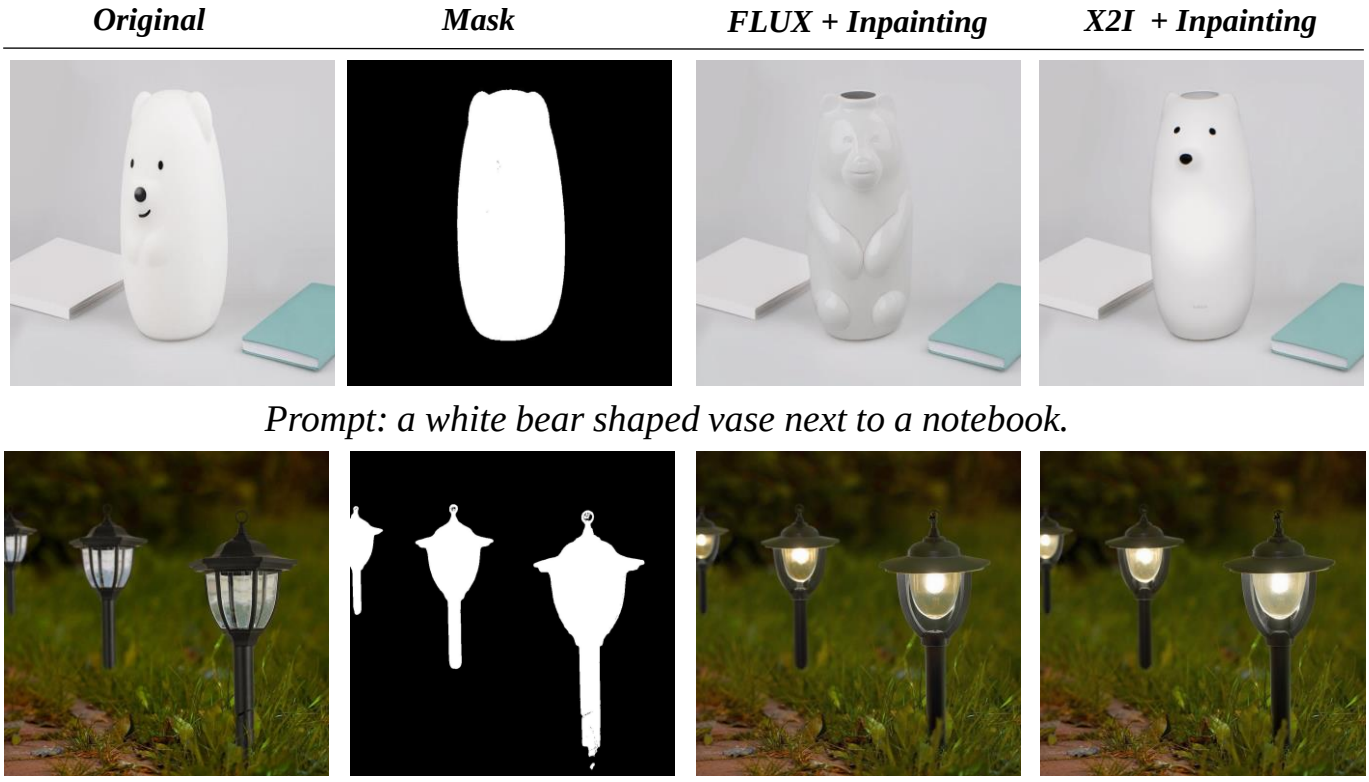


Figure 38. Comparison between FLUX.1 and X2I on the inpainting task.

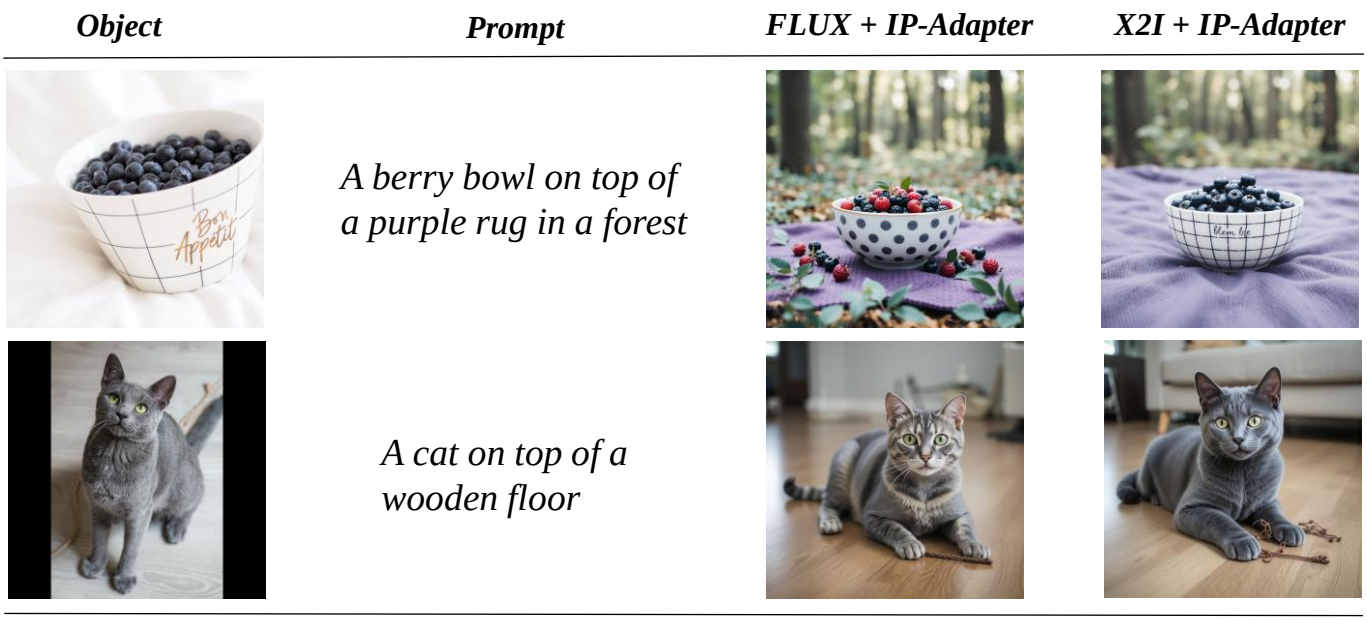


Figure 39. Comparison between Flux.1 and X2I on IP-Adapter downstream tasks.

Figure 40. Comparison between Flux.1 and X2I on ControlNet tasks using canny, depth, and hed as reference maps. The annotation R indicates reference image incorporation in both FLUX.1 and X2I pipelines. Red boxes highlight X2I's superior fine-grained generation.