X-Fusion: Introducing New Modality to Frozen Large Language Models

Sicheng Mo[1] Thao Nguyen[2] Xun Huang[3] Siddharth Srinivasan Iyer[3] Yijun Li[3] Yuchen Liu[3]
Abhishek Tandon[3] Eli Shechtman[3] Krishna Kumar Singh[3] Yong Jae Lee[2] Bolei Zhou[1] Yuheng Li[3]
[1]University of California, Los Angeles [2]University of Wisconsin–Madison [3]Adobe Research
https://sichengmo.github.io/XFusion/

Abstract

We propose X-Fusion, a framework that extends pretrained Large Language Models (LLMs) for multimodal tasks while preserving their language capabilities. X-Fusion employs a dual-tower design with modality-specific weights, keeping the LLM's parameters frozen while integrating vision-specific information for both understanding and generation. Our experiments demonstrate that X-Fusion consistently outperforms alternative architectures on both image-to-text and text-to-image tasks. We find that incorporating understanding-focused data improves generation quality, reducing image data noise enhances overall performance, and feature alignment accelerates convergence for smaller models but has minimal impact on larger ones. Our findings provide valuable insights into building efficient unified multimodal models.

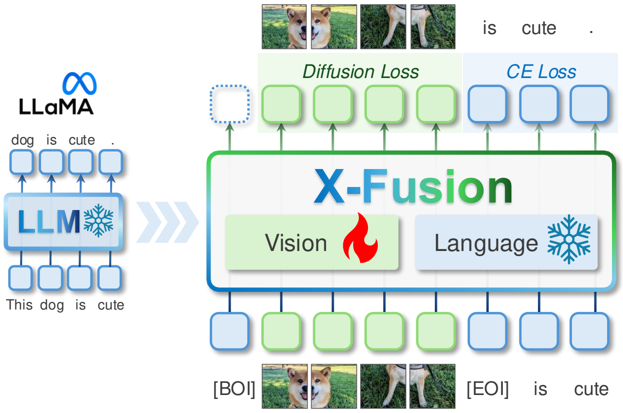


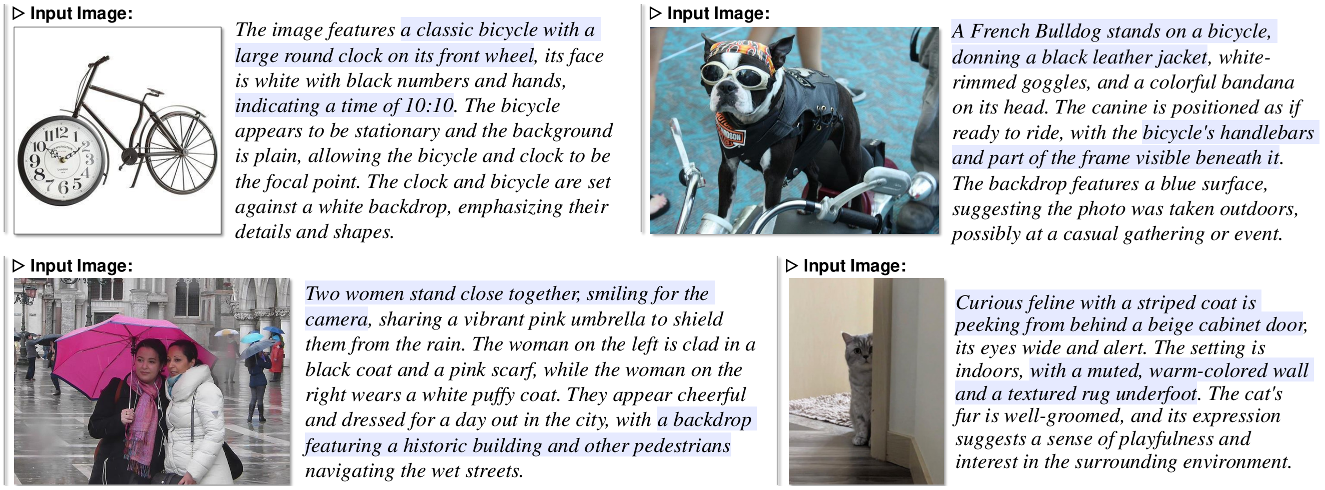
Figure 1. We introduce **X-Fusion** - a novel framework that adapts pretrained LLMs (e.g., LLaMA) to new modalities (e.g., vision) while retaining their language capabilities and world knowledge.

1. Introduction

Large Language Models (LLMs) [1–13] have not only achieved unprecedented capabilities for language processing tasks (e.g., conversational AI [14–20]), but also emerged as foundation tools to solve multiple language-related challenges (e.g., coding [21–23]). However, as humans, we do not communicate solely through text, but also extend to other modalities, such as vision. For example, instead of simply saying "This dog is cute", we might show a photo of the dog to enhance the message: "[image] is cute" (Fig. 1). Thus, a truly versatile AI model should not only understand, reason, and generate textual output, but also must have abilities to understand, reason, and generate visual information. Moreover, these models should be unified to process and generate language and vision simultaneously, creating a more comprehensive interactive experience.

To achieve a unified model, some approaches focus on training unified vision-language models entirely from scratch using next-token prediction loss [24–27]. A recent alternative, Transfusion [28], adopts a domain-specific strategy by combining next-token prediction loss for language with diffusion loss for images. This hybrid architecture has significantly advanced performance, demonstrating greater promise than purely autoregressive approaches like Chameleon [26]. However, training such models from scratch demands immense computational resources (e.g., [28] trained on 2T tokens) and necessitates full retraining for each new modality. Given these shortcomings, another prominent research direction explores how to reuse powerful pretrained LLMs and introduce vision abilities to them [29–31], offering a more practical and efficient way for unified multimodal model training.

Research on adapting LLMs [14–18] with image understanding has shown promising results through "visual instruction tuning" [32–35]. These models typically fine-tune the LLM to align the text feature space with pretrained vision encoders (e.g., CLIP [36]), thus, might degrade original language capabilities [37, 38]. Unlike image understanding, image generation poses greater difficulties, as it demands output capabilities in a new feature space. A large body of work [29, 30, 39–43] tackled this issue by leveraging pretrained image generation models (e.g., Stable Diffusion [44]). However, as these frameworks are not



The image features a classic bicycle with a large round clock on its front wheel, its face is white with black numbers and hands, indicating a time of 10:10. The bicycle appears to be stationary and the background is plain, allowing the bicycle and clock to be the focal point. The clock and bicycle are set against a white backdrop, emphasizing their details and shapes.

A French Bulldog stands on a bicycle, donning a black leather jacket, white-rimmed goggles, and a colorful bandana on its head. The canine is positioned as if ready to ride, with the bicycle's handlebars and part of the frame visible beneath it. The backdrop features a blue surface, suggesting the photo was taken outdoors, possibly at a casual gathering or event.

Two women stand close together, smiling for the camera, sharing a vibrant pink umbrella to shield them from the rain. The woman on the left is clad in a black coat and a pink scarf, while the woman on the right wears a white puffy coat. They appear cheerful and dressed for a day out in the city, with a backdrop featuring a historic building and other pedestrians navigating the wet streets.

Curious feline with a striped coat is peeking from behind a beige cabinet door, its eyes wide and alert. The setting is indoors, with a muted, warm-colored wall and a textured rug underfoot. The cat's fur is well-groomed, and its expression suggests a sense of playfulness and interest in the surrounding environment.

Figure 2. **Captions generated by X-Fusion** demonstrate high details and strong visual alignment with the image inputs.

unified, this approach creates several limitations: limited cross-modal reasoning, restricted in-context learning, and increased error accumulation [45]. Most critically, these approaches typically require fine-tuning the LLM backbone, degrading inherited text generation ability [37, 38]. This raises fundamental research question: *Is there a better way to introduce new modalities to pretrained LLMs?*

Motivated by these observations, we propose **X-Fusion**, a new approach that addresses two challenges: (i) retaining the language abilities of the pre-trained LLM while (ii) adapting it with image generation capabilities. First, X-Fusion freezes all language weights, denoted as the *text tower*, thus preserving the inherent language abilities. Second, instead of fine-tuning the LLM, we introduce a *vision tower* with separate vision weights in each layer to help process visual information for the LLM (Fig 1). This approach aligns text and vision features not only at input or output level, but also at the intermediate processing level. It is worth noting that this architecture is flexible in terms of design — the vision and text towers can have asymmetric architectures. Moreover, the framework naturally extends to additional modalities (e.g., audio) by introducing dedicated modality-specific towers, ensuring efficient and scalable integration while keeping each modality independent.

While architectural innovations are crucial for integrating visual capabilities into LLMs, nowadays, understanding the impact of training data is equally important. Thus, we conduct a comprehensive set of ablation studies from a data-centric perspective to build a scalable training strategy. First, we note that creating image understanding samples without introducing noise to the images is crucial for training diffusion-based unified models. This approach enhances both image generation and understanding abilities due to more semantic visual representation learned from clean images. Based on this key component, we further observed cross task synergy—including more image understanding data that enhances generation performance. We also investigate the effectiveness of aligning our vision features with additional pre-trained representations. Our findings suggest that this extra alignment loss term may help smaller models converge faster, but the benefit diminishes as the model size increases.

In short, our contributions are: (i) X-Fusion - a novel framework that adapts pretrained LLMs to new modalities while retaining their language capabilities. (ii) A systematic study on training strategy, offering insights for optimizing multi-modal learning. (iii) Experimental results on both image-to-text and text-to-image tasks, validating the effectiveness of the proposed architecture.

2. Related Work

Large Multimodal Models. The development of artificial intelligence has historically followed separate, modality-specific paths. Large Language Models (LLMs) [1–13] exclusively processed text input and generated textual responses, while computer vision models specialized in visual content understanding (e.g., object detection [46]) or visual generation (e.g., StyleGAN [47]). The emergence of vision encoder models like CLIP [36] bridged text and image modalities, enabling two key advances: (1) Vision-language models that allow LLMs to “see” (e.g., LLaVA [32]); and (2) Conditional visual generation models that can process textual input (e.g., Stable Diffusion [44]). Current research frontier are focusing on integrating vision and language capabilities into unified models that can both process and generate multimodal content. There are three main approaches: (1) Merging LLMs with pretrained image generation models (e.g., DreamLLM [29], GILL [30]); (2) Training LMMs via next-token prediction (e.g., Chameleon [26]); or (3) Training LMMs using both diffusion and next-token prediction losses (e.g., Transfusion [28]). Following the third approach, which has achieved state-of-the-art results across

A sentient AI robot holding a flower, contemplating human emotions.

A ship sailing across a desert made of stars, its sails catching cosmic winds.

An astronaut riding a camel on Mars during a sunset.

A medieval castle floating upside down over a still lake.

A futuristic warrior queen commanding an army of robotic.

A photo of a whimsical garden where oversized flowers sing in the breeze.

A photo of a cunning fox in a suit and tie reading a newspaper in a city park.

A photo of a red book and a yellow vase on a polished table by a sunny window.

A photo of a playful puppy chasing vibrant holographic bubbles in a park.

A crystalline rose blooming on a frozen lake.

Figure 3. **Images generated by X-Fusion** demonstrate high visual quality and strong text alignment with the input prompts.

modalities, we instead propose initializing from frozen LLMs rather than training from scratch, significantly reducing computational costs and retaining the LLMs knowledge.

Leveraging Pretrained LLMs. Since the success of LLMs [1–13], researchers have discovered that pretrained LLMs can serve as effective baselines for various purposes (e.g., coding [21–23]). Beyond being adapted for domain-specific tasks, prior works have shown that these pretrained models can also be fine-tuned to acquire new abilities, paving the way for a cost-effective approach to include more modalities like image understanding (e.g., LLaVA [32], Mini-GPT-4 [34]), image generation (e.g., GILL [30], DreamLLM [29], SEED-X [40]), or both image understanding and generation (e.g., Show-o [48], Emu3 [27], MetaMorph [31]). However, this method is not without limitations—often, when fine-tuning the LLMs' backbone, it risks compromising the original knowledge. To alleviate this shortcoming, in this work, we propose a novel method to extend vision-related abilities for LLMs while keeping all the language layers frozen, thus maintaining its original language abilities untouched. A concurrent work is LMFusion [49], with a similar high-level approach. However, they use joint attention across text and vision tokens, which is less flexible than our proposed model design.

Task-specific Weights. People have been exploring the use of specialized models tailored to different tasks in both the language [50–53] and vision domains [54–56]. The use of unshared parameters has also been explored in multi-modal settings. For instance, CLIP [36] and ImageBind [57] focus on representation learning, while other works [58–60] emphasize vision-conditioned language model pretraining. However, these multi-modal approaches predominantly focus on visual understanding, neglecting generation tasks. Importantly, they are typically trained from scratch, which is computationally expensive. A concurrent work, Playground-v3 [61], takes a different approach by building upon LLMs, but freezes them to perform generation tasks only, without addressing visual understanding.

3. Preliminaries

In this section, we provide a brief preliminary overview of the state-of-the-art recipe (proposed by [28]) for training unified models in a hybrid manner, incorporating next-token prediction for language and diffusion models for image.

3.1. Language Modeling via Autoregression

LLMs are typically trained using an autoregressive modeling objective, where the joint probability of a sequence of language tokens $\mathbf{x}^{\text{txt}} = \{x_1^{\text{txt}}, x_2^{\text{txt}}, \ldots, x_N^{\text{txt}}\}$ is factorized as a product of conditional probabilities:

$$P_\theta(\mathbf{x}^{\text{txt}} \mid c) = \prod_{i=1}^{N} P_\theta(x_i^{\text{txt}} \mid x_{<i}^{\text{txt}}, c).$$

Here, c represents optional conditioning information, which could include extracted features from other modalities (e.g., image representations) or task-specific context. However, in the standard setting, LLMs are usually trained without any additional conditioning (c is absent), and the predictions depend solely on the preceding tokens $x_{<i}^{\text{txt}}$.

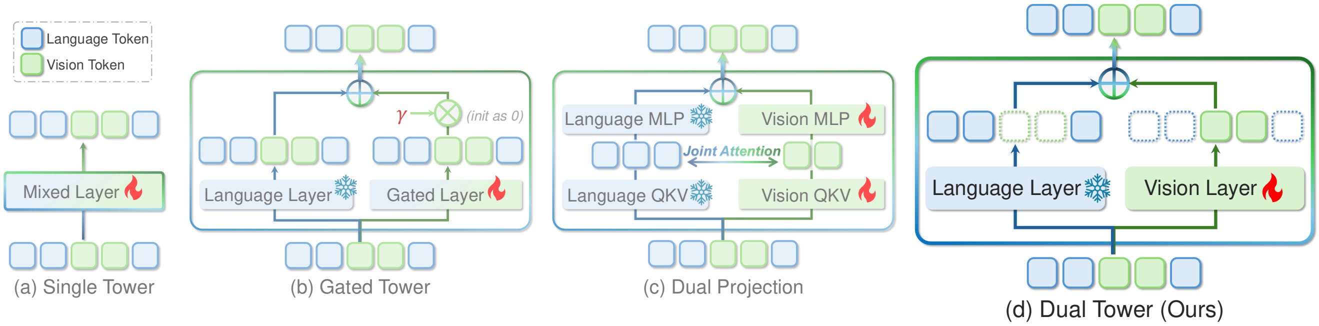


Figure 4. **Conceptual comparison of four model architecture baselines.** Here, we illustrate how each layer processes the sequential multi-modal feature. **(a) Single Tower:** Directly fine-tuning pre-trained LLM. **(b) Gated Layer:** Duplicate Each LLM layer as the gated vision layer, **(c) Dual Projection:** Duplicate QKV matric and MLP layer for vision modality, **(d) Dual Tower:** Duplicated transformer block for vision modality.

The training objective for this autoregressive model is to minimize the negative log-likelihood of the observed data:

$$\mathcal{L}_{\text{AR}} = \mathbb{E}_{\mathbf{x}^{\text{txt}}, c} \left[- \sum_{i=1}^{N} \log P_\theta(x_i^{\text{txt}} \mid x_{<i}^{\text{txt}}, c) \right].$$

If conditioning information c is introduced (e.g., in multimodal or task-specific setups), it allows the model to extend its capabilities to tasks such as conditional text generation and cross-modal understanding.

3.2. Image Modeling with Diffusion

Diffusion models have emerged as one of the most successful approaches for image generation thanks to its ability to generate high-quality images. Diffusion models are trained to gradually reverse the process of adding noise to an image, starting from a noise vector $\mathbf{x}_T^{\text{img}}$ and progressively generating less noisy samples $\mathbf{x}_{T-1}^{\text{img}}, \mathbf{x}_{T-2}^{\text{img}}, \ldots, \mathbf{x}_0^{\text{img}}$. The goal is to produce high-quality images by learning a denoising function f_θ parameterized by θ that can reverse this process.

The diffusion model training objective aims to minimize the difference between the predicted and true noise. Specifically, for each time step t, the objective is to solve the following denoising problem on the image data $\mathbf{x}^{\text{img}}$:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{\mathbf{x}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \left[\| \epsilon - f_\theta(\mathbf{x}_t^{\text{img}}, t, \mathbf{c}) \|_2^2 \right],$$

where $\mathbf{x}_t^{\text{img}}$ is the noisy image at time step t, uniformly sampled from $\{1, \ldots, T\}$, and $f_\theta(\mathbf{x}_t^{\text{img}}, t, \mathbf{c})$ is the denoising function that predicts the noise added to $\mathbf{x}_t^{\text{img}}$ conditioned on the time step t and context $\mathbf{c}$ (often is text prompt).

4. X-Fusion

In this section, we describe X-Fusion, an unified framework that adapts pretrained LLMs for vision tasks while preserving their inherent language capabilities. As illustrated in Fig. 1, X-Fusion processes both image and text inputs within a single model. To maintain the pretrained LLM's language knowledge, we freeze its weights and introduce new trainable parameters to handle vision inputs.

Tokenizer. Text inputs are tokenized using the original LLM's tokenizer to produce text tokens. Images are processed through a pretrained visual encoder to obtain latent representations. The encoders can be either a low-level feature compression model, like latent VAE [62], or CLIP [36]/DINO [63]-like semantic encoders. These latent representations can optionally be further encoded into vision tokens using an additional trainable vision layer. In most of our experiments, we use VAE from SD1.5 [62] as the pretrained visual encoder. After that, we use a trainable linear patchify layer with a patch size of 2 × 2 to further compress image features into image tokens. The combined interleaved image-text tokens are then passed into X-Fusion, where the trainable parameters facilitate joint optimization for both image understanding and generation.

Dual Tower. Let denote the tokenized input embeddings as:

$$\mathbf{E}^{\text{in}} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M\}.$$

where token $\mathbf{e}_i$ can be either text or vision token.

To effectively process a mixture of textual and visual modalities while retaining the original language information, each layer in X-Fusion should incorporate new trainable weights alongside a frozen language layer. Specifically, we introduce two components for each layer: A frozen text transformer block $\mathcal{F}^{\text{txt}}$, and a trainable vision transformer block $\mathcal{F}^{\text{img}}$.

Both components operate on the embeddings $\mathbf{E}^{\text{in}}$, and their respective outputs are given by:

$$\mathbf{H}^{\text{txt}} = \mathcal{F}^{\text{txt}}(\mathbf{E}^{\text{in}}), \quad \mathbf{H}^{\text{img}} = \mathcal{F}^{\text{img}}(\mathbf{E}^{\text{in}}).$$

Intuitively, the vision layer needs to process visual tokens conditioned on texture features for generation tasks. Additionally, it also need to extract features suitable for visual understanding tasks, ensuring that the visual features can be effectively interpreted by the frozen language layer.

The outputs of the text and vision blocks are selectively combined to form the output sequence of the block:

$$\mathbf{H}^{\text{out}} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_M\},$$

where

$$\mathbf{h}_i = \begin{cases} \mathbf{h}_i^{\text{txt}}, & \text{if } x_i \in \mathbf{x}^{\text{txt}}, \\ \mathbf{h}_i^{\text{img}}, & \text{if } x_i \in \mathbf{x}^{\text{img}}. \end{cases}$$

Here, $\mathbf{h}_i^{\text{txt}} \in \mathbf{H}^{\text{txt}}$ and $\mathbf{h}_i^{\text{img}} \in \mathbf{H}^{\text{img}}$ are the outputs of the text and vision blocks, respectively, for the i-th token. In our design, we initialize each vision block by copying the parameters in the corresponding language transformed layer, which consists of one self-attention layer, one MLP layer, and two normalization layers.

Finally, each text embedding is decoded into discrete tokens through a linear classification head. The output image feature modeling can be flexible, and could involve methods such as diffusion (L2-loss), continuous autoregressive modeling (cosine regression), or discretized autoregressive modeling (cross-entropy), depending on the choice of pretrained visual encoder and design decisions. In this paper, we primarily focus on diffusion modeling. It is important to note that the best modeling approach for new modalities is case-specific and may be explored as future work.

X-Fuse (Optional). In the explanation above, we select vision tokens from the vision layers and text tokens from the text layers. Thus, in practice, we do not need to calculate the vision query tokens in the text layer, nor the text query tokens in the vision layers. This is a main design choice, as it results in the same attention FLOPs as other baseline variants shown in Fig. 4, which we will discuss in the experimental section.

Optionally, we introduce an operation called X-Fuse, which sacrifices FLOPs to improve performance. Taking the text feature as an example, we also calculate the text query features in the vision layer. To fuse the text feature from both towers, we compute:

$$\alpha * \mathbf{h}_i^{\text{txt-txt}} + \beta * \mathbf{h}_i^{\text{txt-img}}$$

Here, the superscript indicates the source tower of the text features: ”txt-txt” refers to text features from the text tower, and ”txt-img” refers to text features from the vision tower. The scalars α and β are learnable parameters. A similar operation can be used to fuse image features as well.

Training. The final training objective combines the autoregressive loss ($\mathcal{L}_{\text{AR}}$) and the image denoising loss ($\mathcal{L}_{\text{DM}}$), with their respective weighting coefficients λ_{AR} and λ_{DM}:

$$\mathcal{L} = \lambda_{\text{AR}} \cdot \mathcal{L}_{\text{AR}} + \lambda_{\text{DM}} \cdot \mathcal{L}_{\text{DM}}$$

We choose LLaMA-3 family [17] as our as pretrained LLMs for X-Fusion and use the flow matching scheduler following Stable Diffusion 3 [64]. We set $\lambda_{\text{AR}} = 0.2$ and $\lambda_{\text{DM}} = 1$, then train with AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$), a linear warm-up scheduler, a maximum learning rate of $lr = 1 \times e^{-4}$, and DeepSpeed Stage 2 [65] distributed training on H100 GPUs. We train with batches of 0.8M tokens for 100k steps, processing a total of 0.08T tokens for most experiments. We extend the training to 200K steps on our X-Fusion model initialized from LLaMA-3.1-8B and report the evaluation results in Supplementary.

	text only	text2img		img2text
Model	**MMLU (↑)**	**FID (↓)**	**CLIP (↑)**	**BLIP(↑)**
LLaMA3.2-1B [17]	32.2	—	—	—
Single Tower	25.0	19.10	22.63	30.2
Gated Tower	**32.2**	24.51	21.91	14.5
Dual Projection	**32.2**	20.22	22.46	30.9
*Dual Tower (**Ours**)*	**32.2**	**14.20**	**22.81**	**31.3**

Table 1. **Architecture design comparison.** Our dual-tower approach surpasses other baselines in image generation tasks and delivers competitive performance in image understanding, while maintaining the original language capability.

Task. We evaluate X-Fusion's performance on: (i) image understanding and (ii) image generation, as these tasks reflect the connection between visual and textual modalities.

- *Image Generation*: To assess image generation, we evaluate text-to-image performance. Specifically, we generate 30K images using randomly sampled prompts from the MS-COCO [66] and report image quality using FID [67].
- *Image Understanding*: To assess image understanding, we evaluate image captioning (or image-to-text) performance. Specifically, we generate captions for 30K images from the MS-COCO dataset [66] and report caption quality using BLIP2-ITM [68]. We also considered additional metrics, such as CIDEr [69] and BertScore [70], however, these were either inappropriate for long captions or failed to capture meaningful changes during training. Further analysis can be found in Supplementary.

Data. Otherwise stated, we sample 0.08T tokens/patches from an in-house licensed dataset. We create image-caption pairs by center-cropping and resizing images to 256 × 256, and pairing them with detailed captions generated by the InternVL-2.0 26B model [35]. These pairs are then formatted for both image-to-text tasks (serving as understanding data), and text-to-image tasks (serving as generation data).

5. Architecture Design Choices for Adding Vision Abilities

To evaluate the effectiveness of our Dual Tower design, we compare it against three alternative transformer block variants (Fig. 4) that are designed for multimodal integration. Apart from the transformer blocks, all other components—including tokenizers, encoder, and decoder modules—are kept identical across configurations.

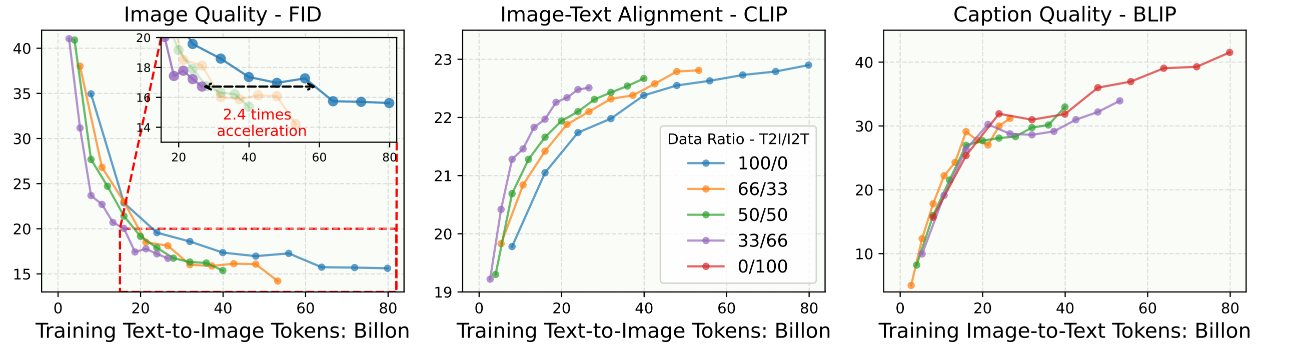


Figure 5. **Performance of image generation and understanding at various data ratios.** Increasing visual understanding data improves visual generation performance.

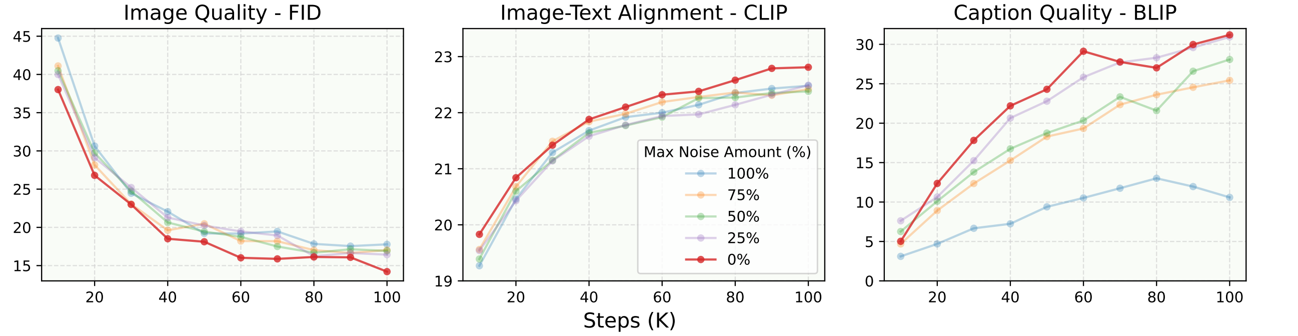


Figure 6. **Performance of image generation and understanding at various noise limits in the image-to-text samples.** Providing clear images for image-to-text samples enhances visual generation and understanding simultaneously.

- *Single Tower:* This simple baseline uses the original language model transformer block to process both inputs directly. Note that this is equivalent to Transfusion [28] but with pre-trained LLM as initialization (Fig. 4a).
- *Gated Tower:* Inspired by [71], we duplicate the language transformer block into a trainable "gated block" (Fig. 4b), with both blocks taking the same input sequence:

  $$\mathbf{H}^{\text{txt}} = \mathcal{L}^{\text{txt}}(\mathbf{E}^{\text{in}}), \quad \mathbf{H}^{\text{gate}} = \mathcal{L}^{\text{gate}}(\mathbf{E}^{\text{in}}).$$

  After that, the gated block output will be added to the language transformer block output, multiplied by a learnable value γ which is initialized as 0:

  $$\mathbf{H}^{\text{out}} = \mathbf{H}^{\text{txt}} + \gamma * \mathbf{H}^{\text{gate}}.$$

- *Dual Projection:* We copy the language weights consisting of one self-attention and one MLP (Fig. 4c). However, the data flow is different from dual-tower. First of all, based on the modality of $\mathbf{e}_i$, we apply *separate projection matrices* for query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$), and then a joint attention is operated on all tokens:

  $$\mathbf{H}^{\text{attn}} = \text{attn}(\text{QKV}_{\text{modality}}(\mathbf{e}_i)),$$

  After that, the output feature from self-attention is passed through a modality-specific MLP:

  $$\mathbf{H}^{\text{out}} = \text{MLP}_{\text{modality}}(\mathbf{H}^{\text{attn}}).$$

  This variant is similar to concurrent work [49]. However, dual-tower offers more flexibility, as the vision and language layers can be designed differently.

FLOPs. Let N denote the number of text tokens and M denote the number of image tokens. For all three alternative designs, the attention complexity is $O((N+M)^2)$. In our dual-tower design, for a fair comparison, we choose not to use the X-Fuse operation. As a result, the complexity in the vision tower is $O(M \cdot (N+M))$ (since we do not need query tokens for text), and the complexity in the text tower is $O(N \cdot (N+M))$ (since we do not need query tokens for image). In total, this results in the same complexity of $O((N+M)^2)$.

Quantitative Results. Tab. 1 presents a comparison of different architectural designs, all are using pretrained LLaMA-3.2-1B [17]. Among them, the Dual Tower architecture achieves the best performance in both image generation and understanding tasks. In contrast, the Gated Layer architecture performs the weakest in both tasks, likely due to the limitations of simple addition operations. Among baselines, the Single-Tower model delivers decent performance across both tasks; however, our Dual Tower model achieves a 23% lower FID while maintaining the same number of training parameters. More importantly, the Single-Tower model compromises the inherent knowledge of the original language model due to training on T2I and I2T tasks. To assess the model's general knowledge, we evaluate it on the MMLU benchmark [72], a multiple-choice test with four answer options, using a 5-shot setting. Results show that the Single Tower model's performance drops to 25.0, which is equivalent to chance-level performance.

Dual Tower and Dual Projection share a common insight: modality-specific operations. While their text generation capabilities are nearly same, Dual Tower outperforms in image generation. This superiority can be attributed to its flexibility: The vision layers in Dual Tower can generate new QKV representations for input text features, followed by attention operations between the image and the new text features. In contrast, Dual Projection is limited to using the original language model's text QKV matrices for attention computation. Also, it is worth noting that conceptually, Dual Tower offers greater flexibility: while our current implementation replicates the language layer as the vision layer, the vision layer does not have to be identical to the language layer, as long as it produces outputs with the same feature dimensions. *For the rest of the paper, all experiments will use Dual Tower design*.

> Text processing flexibility affects image generation but not image understanding performance.

6. Effect of Data Ratios and Noise on Generation and Understanding Tasks

We now turn to examine the effects of training data on X-Fusion's performance. In the following section, we address two key questions: (i) How does noise level affect performance in a joint training setting? (ii) Does multitask training provide mutual benefits between tasks?

6.1. Effect of Noise Amount

As a unified model, it should be able to perform both image denoising and text autoregressive tasks simultaneously on input data. However, the level of noise applied to images in image-to-text (I2T) samples remains a question. While diffusion-based image modeling benefits from noisy input for generation tasks, excessive noise can degrade visual quality and hinder image understanding. This issue was also noted by [28], where they proposed limiting the diffusion noise on I2T samples to a maximum of $t = 50\%T$ to reduce distortion for visual understanding while still can treat those noisy images as generation training data.

We argue that while this approach helps, it may not be optimal. We hypothesize that training with clean images (i.e., without adding noise to I2T samples) can lead to a stronger vision tower for image understanding, ultimately improving generation quality, although this reduces the amount of denoising data available for generation.

To validate our hypothesis,we conducted comprehensive experiments where we systematically varied the max noise level for image understanding (I2T) tasks: 0% (clean image), 25%, 50%, 75%, and 100% (normal generation setting). Throughout these experiments, we maintained a 1:2 data ratio between understanding and generation tasks. Results are provided in Fig. 6. As can be seen, generally the more noisy the images are, the more image understanding performance is degraded. Our proposed strategy (providing clean image for understanding) consistently achieves the best results for image understanding (2nd and 3rd col.). Interestingly, using clean image for understanding also helps to boost performance of image generation! (1st col.).

> Using clean images for visual understanding improve performance for both tasks.

6.2. Effect of Data Ratio

Along with the noise addition strategy, task data ratio is also a major factor for training unified multimodal models. To investigate the synergy between visual understanding and generation from data's perspective, we kept the architecture unchanged and trained for 100k steps with a batch size of 0.8M tokens on a total of 0.08T tokens, varying the composition of text-to-image (T2I) and image-to-text (I2T) tasks. Specifically, we begin with the training data composed entirely of T2I tasks (100% T2I, 0% I2T, or denoted as 100/0 for short), then progressively decrease the proportion of T2I data while increasing I2T data (i.e., 66/33, 50/50, 33/66) until the dataset consists solely of I2T tasks (0/100).

Fig. 5 illustrates the results. To investigate how training data from one task influences performance on other tasks, we plot performance metrics against the number of tokens observed for each specific task (e.g., first column, x-axis indicates how many image generation tokens are trained). In another words, we want to ask: When two models are trained on an identical number of image generation tokens, but differ in their exposure to image understanding tokens, how do their performance diverge?, and vice versa.

We observe that incorporating image understanding data (I2T) improves generation quality (T2I). As the proportion of I2T data increases while keeping total T2I data fixed, generation performance consistently improves (1st and 2nd panels). In contrast, visual generation data (T2I) does not positively impact the understanding task (I2T) (third panel). Overall, there's an asymmetric relationship: understanding data benefits generation, but generation data does not enhance understanding. Based on our findings, we recommend a 66/33 (or 2:1) training ratio, which strikes a strong balance and optimizes performance across both tasks.

> Incorporating image understanding data enhances image generation performance, while adding generation data does not impact image understanding tasks.

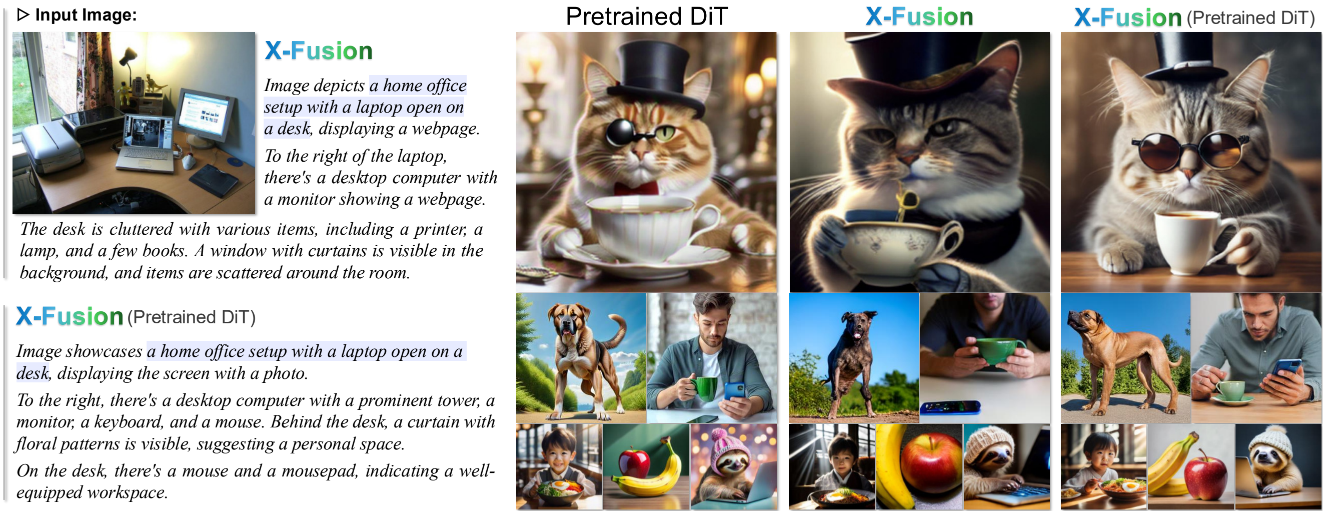


Image depicts a home office setup with a laptop open on a desk, displaying a webpage. To the right of the laptop, there's a desktop computer with a monitor showing a webpage. The desk is cluttered with various items, including a printer, a lamp, and a few books. A window with curtains is visible in the background, and items are scattered around the room.

Image showcases a home office setup with a laptop open on a desk, displaying the screen with a photo.

To the right, there's a desktop computer with a prominent tower, a monitor, a keyboard, and a mouse. Behind the desk, a curtain with floral patterns is visible, suggesting a personal space.

On the desk, there's a mouse and a mousepad, indicating a well-equipped workspace.

Figure 7. **Qualitative comparison** between pretrained DiT model, X-Fusion(DiT) and vanilla X-Fusion on image generation and understanding task. By initializing the vision tower from pretrained text-to-image diffusion model, X-Fusion(DiT) achieves stronger image generation capability and competitive performance on image understanding compared to vanilla X-Fusion.

7. Extension of X-Fusion

In this section, we introduce three extensions of X-Fusion, including X-Fuse layer, transferring from pretraind diffusion models, and finetune for downstream tasks.

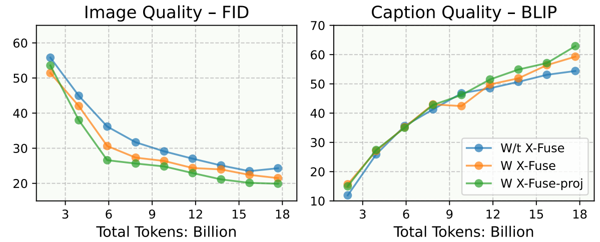


Figure 8. **Ablation: X-Fuse layer**. Our X-Fusion model with the X-Fuse layer outperforms the baseline X-Fusion model on image generation and understanding tasks.

7.1. Effectiveness of X-Fuse Layer

So far, our study is conducted on a dual-tower architecture without the X-Fuse operation, maintaining the same FLOPs as other baseline designs. In Sec 5, we further propose the X-Fuse operation, which merges features from both towers to trade increased FLOPs for improved performance. We conduct an ablation study to evaluate this design, and the results are shown in Fig 8. As illustrated, applying the X-Fuse operation leads to improvements in both capabilities.

7.2. Transfer from Pretrained Diffusion Model

While X-Fusion successfully kept its language generation capability, its image generation capability still needs to be trained from scratch. One solution is to transfer the knowledge from existing image generation models. In our dual-tower design, each block in the language and vision tower processes the entire feature sequence independently, therefore allowing non-identical block designs in both towers. With this advantage, we could transfer the image generation capability from a large-scale pretrained diffusion model that uses diffusion transformers [64, 73].

We train a variation of X-Fusion using Llama3.1-8B as the language tower and an in-house pretrained text-to-image DiT model as its vision tower, notated as X-Fusion(Pretrained DiT), using the same training recipe in the previous section for 50K iterations. To align the feature dimension in both towers, we add linear projection layers within the X-Fuse layer. Figure 8 shows that this operation further enhances the model's capability. Figure 7 qualitatively compares the image generation and image understanding performance between the pre-trained DiT model, X-Fusion(Pretrained DiT), and vanilla X-Fusion-8B models. X-Fusion(Pretrained DiT) obtained stronger image generation capability and on-par image understanding performance compared to the vanilla X-Fusion.

8. Conclusion

This paper introduces X-Fusion, a novel framework for adapting pretrained Large Language Models to new modalities (e.g., vision) while retaining their original language capabilities. We propose a Dual Tower architecture in which language weights remain frozen, while visual features are processed via a trainable vision tower with separate weights. Alongside this novel architecture, we provide a systematically comprehensive set of ablation studies that offer valuable insights from a data perspective. Our findings reveal that: (i) incorporating understanding-focused data improves generation performance, (ii) reducing noise in image data enhances overall results, and (iii) feature alignment benefits primarily smaller models. We hope our paper will step forward building an unified Large Multimodal Models in a more effiecient way.

Acknowledgment: The work was support in part by Amazon Research Award through the UCLA Science Hub, NSF IIS2404180, Microsoft Accelerate Foundation Models Research Program, and Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training).

References

[1] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 1, 2, 3

[2] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[3] Tom B. Brown and et al. Language models are few-shot learners, 2020.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

[5] Jack W. Rae and et al. Scaling language models: Methods, analysis & insights from training gopher, 2022.

[6] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022.

[7] Jordan Hoffmann and et al. Training compute-optimal large language models, 2022.

[8] PaLM team. Palm: Scaling language modeling with pathways, 2022.

[9] PaLM 2 team. Palm 2 technical report, 2023.

[10] Albert Q. Jiang and et al. Mixtral of experts, 2024.

[11] Albert Q. Jiang and et al. Mistral 7b, 2023.

[12] DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism, 2024.

[13] Gemma Team. Gemma: Open models based on gemini research and technology, 2024. 1, 2, 3

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 1

[15] Llama 2 team. Llama 2: Open foundation and fine-tuned chat models, 2023.

[16] Phi-3 team. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[17] Abhimanyu Dubey and et al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024. 5, 6

[18] Qwen team. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[19] InternLM2 team. Internlm2 technical report, 2024.

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

[21] Mark Chen and et al. Evaluating large language models trained on code, 2021. 1, 3

[22] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis, 2023.

[23] Raymond Li and et al. Starcoder: may the source be with you!, 2023. 1, 3

[24] Lili Yu and et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023. 1

[25] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.

[26] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 2

[27] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 1, 3

[28] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. 2024. 1, 2, 3, 6, 7

[29] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation, 2024. 1, 2, 3

[30] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023. 1, 2, 3

[31] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 1, 3

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[34] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[35] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 5

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 4

[37] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning (Proceedings Track)*, 2023. 1, 2

[38] Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. Model tailor: mitigating catastrophic forgetting in multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. 1, 2

[39] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023. 1

[40] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 3

[41] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[42] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context, interleaved, and interactive any-to-any generation. 2023.

[43] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 1

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[45] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017. 2

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 2

[47] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 2

[48] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 3

[49] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 3, 6

[50] Noam M. Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ArXiv*, abs/1701.06538, 2017. 3

[51] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam M. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *ArXiv*, abs/2006.16668, 2020.

[52] William Fedus, Barret Zoph, and Noam M. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *ArXiv*, abs/2101.03961, 2021.

[53] Nan Du and et al. Glam: Efficient scaling of language models with mixture-of-experts. 2021. 3

[54] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Neural Information Processing Systems*, 2021. 3

[55] Yuheng Li, Yijun Li, Jingwan Lu, Eli Shechtman, Yong Jae Lee, and Krishna Kumar Singh. Collaging class-specific gans for semantic image synthesis. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14398–14407, 2021.

[56] Jiapeng Zhu, Ceyuan Yang, Kecheng Zheng, Yinghao Xu, Zifan Shi, and Yujun Shen. Exploring sparse moe in gans for text-conditioned image synthesis. *ArXiv*, abs/2309.03904, 2023. 3

[57] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 3

[58] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv*, abs/2208.10442, 2022. 3

[59] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv*, abs/2111.02358, 2021.

[60] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *ArXiv*, abs/2303.07226, 2023. 3

[61] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *ArXiv*, abs/2409.10695, 2024. 3

[62] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 4

[63] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao

Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 4

[64] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 5, 8

[65] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. 5

[66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[67] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[68] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5

[69] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 5

[70] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. 5

[71] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 6

[72] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. 7

[73] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 8