

# EasyGen: Easing Multimodal Generation with BiDiffuser and LLMs

Xiangyu Zhao, Bo Liu\*, Qijiong Liu\*, Guangyuan Shi\*, Xiao-Ming Wu✉

Department of Computing, The Hong Kong Polytechnic University  
{xiang-yu.zhao, bokelvin.liu, jyonn.liu, guang-yuan.shi}@connect.polyu.hk,  
xiao-ming.wu@polyu.edu.hk

## Abstract

We present EasyGen, an efficient model designed to enhance multimodal understanding and generation by harnessing the capabilities of diffusion models and large language models (LLMs). Unlike existing multimodal models that predominately depend on encoders like CLIP or ImageBind and need ample amounts of training data to bridge modalities, EasyGen leverages BiDiffuser, a bidirectional conditional diffusion model, to foster more efficient modality interactions. EasyGen achieves text generation by training a projection layer linking BiDiffuser and an LLM, and facilitates image generation by training an adapter to align the LLM’s text space with the BiDiffuser’s image space. Comprehensive quantitative and qualitative experiments show that EasyGen excels in data-efficient training, high-quality image generation, and extendibility, effectively addressing the challenges in multimodal generation. The source code is available at <https://github.com/zxy556677/EasyGen>.

## 1 Introduction

In recent years, remarkable progress has been made in the field of artificial intelligence generated content (AIGC), notably in technologies like large language models (LLMs) (Chiang et al., 2023; Touvron et al., 2023; Brown et al., 2020; Chowdhery et al., 2022; Zeng et al., 2022) for text generation and diffusion models (Rombach et al., 2022; Nichol et al., 2022; Saharia et al., 2022) for visual generation. These breakthroughs have paved the way for the development of multimodal large language models (MLLMs), sparking a recent trend of incorporating extra visual modules into LLMs. Collaborative models, such as Visual ChatGPT (Wu et al., 2023a) and MM-REACT (Yang et al., 2023), strategically use externally pre-trained tools to translate visual information into text descriptions and feed

the data into LLMs. However, they are exclusively dependent on pre-trained tools for inference. Contrarily, end-to-end trained models including the BLIP series (Li et al., 2023b), LLaVA series (Liu et al., 2023b,a), MiniGPT-4 (Zhu et al., 2023), and mPLUG-Owl (Ye et al., 2023) focus on mapping image information to the text space of LLMs, enabling LLMs to comprehend visual inputs.

Existing end-to-end models are also not without limitations. First, most of these multimodal models rely on either CLIP (Radford et al., 2021) or ImageBind (Girdhar et al., 2023) as their image encoder. While these encoders excel in learning unified representations that encompass both text and images, they face challenges when it comes to transforming between different modalities. This predicament makes current vision-language models relying heavily on sizable data sets to align CLIP/Bind-encoded images with the language model, due to the disparity between different modalities.

Moreover, the majority of previous multimodal models have concentrated on comprehending multimodal content and lacked the capability to generate multimodal responses, such as content beyond text. Several concurrent works, such as Emu (Sun et al., 2023) and NExT-GPT (Wu et al., 2023), have utilized diffusion models for multimodal generation. Typically, these methods involve training a projection layer to align the output embedding space of the LLM with the input embedding space of the diffusion model (encoded by CLIP’s text encoder) using an MSE loss. However, this approach may lead to the underutilization of the semantic understanding and reasoning capabilities of the LLM, and may introduce information loss in the alignment process, ultimately leading to lower image generation quality compared to the original diffusion model, as elaborated in Sec. 5.6 and Tab. 6.

In this work, we propose EasyGen, an end-to-end model that facilitates multimodal generation with a single bidirectional conditional diffusion

\*Co-second author.

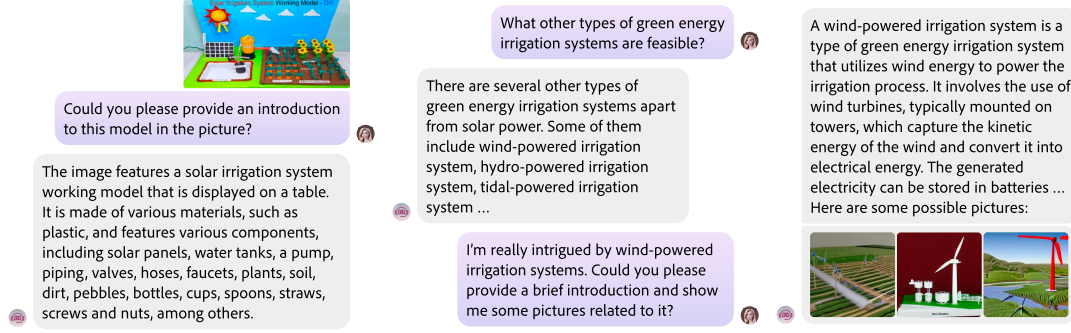


Figure 1: Our model EasyGen can understand multimodal inputs and generate multimodal responses, as illustrated by model-generated speech bubbles in grey color, which include both text and images.

model and LLMs, as illustrated in Figure 2. The diffusion model, called BiDiffuser, is obtained by fine-tuning the UniDiffuser (Bao et al., 2023b), with a specific focus on targeted image-to-text and text-to-image tasks. This fine-tuning addresses UniDiffuser’s limitation of attempting to fit all conditional distributions, including those based on noisy inputs, into a single model, which reduces its effectiveness on specific tasks like conditional generation from noise-free inputs. BiDiffuser plays a pivotal role for both text and image generation. In EasyGen, text generation is achieved by training a projection layer that connects BiDiffuser and an LLM, while image generation is facilitated by training an adapter that infuses the text representation of the LLM into BiDiffuser. Figure 1 showcases EasyGen’s ability to handle multimodal inputs and generate appropriate multimodal responses.

EasyGen holds three significant advantages that address the challenges in multimodal generation:

First, EasyGen offers competitive performance in a *data-efficient* way compared to cutting-edge models, as shown in **Tab. 3 (Sec. 5.5)**. This is due to BiDiffuser’s ability to simplify the alignment of its embedding space with an LLM, which allows for efficient training with less data for image-to-text tasks such as image captioning and VQA.

Second, EasyGen exhibits *superior image generation quality*, surpassing other end-to-end MLLMs, as shown in **Tab. 6 (Sec. 5.6)**. This is attributed to the adapter’s design (Sec. 4.2), which aligns the LLM’s text space with the diffusion model’s image space, thereby utilizing the LLM’s semantic understanding and reasoning capabilities. In contrast, the projection layers in other MLLMs like NExT-GPT only align the LLM’s text space with the diffusion model’s text space and are not trained by the image denoising objective.

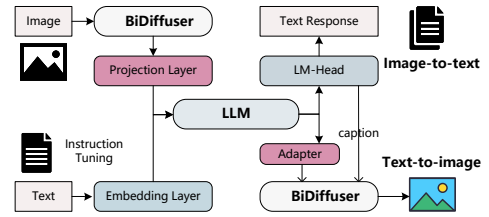


Figure 2: Overview of EasyGen.

Third, EasyGen can be *readily adapted* to manage complex vision-language tasks by incorporating more advanced visual encoders or by integrating BiDiffuser into contemporary sophisticated multimodal LLMs like LLaVA to enhance performance, as shown in **Tab. 5 (Sec. 5.7)**.

## 2 Related Work

**Multimodal Language Models.** Recent research has witnessed a surge of interest in multimodal LLMs, including collaborative models (Wu et al., 2023a; Yang et al., 2023; Shen et al., 2023) and end-to-end methods (Alayrac et al., 2022; Guo et al., 2022; Li et al., 2022; Bao et al., 2021; Wang et al., 2022b,a). More recently, some works also explore training LLMs with parameter-efficient tuning (Li et al., 2023b; Zhang et al., 2023a) and instruction tuning (Dai et al., 2023; Liu et al., 2023b; Ye et al., 2023; Zhu et al., 2023; Li et al., 2023a). Different from them, EasyGen is built upon BiDiffuser, which promotes more efficient interactions between modalities.

**Multimodal Diffusion Models.** Diffusion generative models (Rombach et al., 2022; Ramesh et al., 2021; Nichol et al., 2022; Ruiz et al., 2023) have achieved strong results in text conditioned image

generation works. Specifically, Versatile Diffusion (Xu et al., 2023) employs the U-Net (Ronneberger et al., 2015) architecture with a multi-flow design to tackle multiple modalities and tasks, while UniDiffuser (Bao et al., 2023b) adopts the U-ViT (Bao et al., 2023a) framework to treat both image and text as sequential token streams for diffusion calculations. However, these models are unable to complete complex language tasks. EasyGen combines the advantages of diffusion models and LLMs and achieves competitive performance in both image-to-text and text-to-image tasks.

**Multimodal Response Generation.** Recent research has made significant advancements in multimodal response generation (Koh et al., 2023b; Tang et al., 2023; Zhang et al., 2023b; Wu et al., 2023b; Pan et al., 2023; Koh et al., 2023a; Sun et al., 2023; Dong et al., 2023) using text-to-image models such as Stable Diffusion. However, the lack of semantic understanding capability in the CLIP text encoder may result in low-quality generated images. EasyGen addresses this issue by transferring knowledge from LLM to BiDiffuser via an adapter, enabling the creation of high-quality textual semantic representations for text-to-image generation.

### 3 Basics of Diffusion Models

**Unconditional Generation.** Given a data sample taken from a real data distribution  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ , diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) first destruct the data by constructing a Markov forward process and gradually injecting noise to the data:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad (1)$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$$

where  $\beta_t \in (0, 1)$  is the variance added at diffusion step  $t$ . Then, they learn to reverse the process:

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad (2)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_t(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}),$$

where  $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; 0, \mathbf{I})$  is the standard Gaussian distribution and  $\mu_t(\cdot)$  is the parameterization of the predicted mean. Diffusion models are trained to maximize the marginal likelihood of the data  $\mathbb{E}[\log p_\theta(\mathbf{x}_0)]$ , and the canonical objective is the variational lower bound of  $\log p_\theta(\mathbf{x}_0)$ . Denoising

diffusion probabilistic models (Ho et al., 2020) generate samples  $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$  by injecting noise  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$  to the data  $\mathbf{x}_0$ , and train a network  $\epsilon_\theta(\cdot)$  to predict the added noise  $\epsilon$  using a standard mean squared error loss:

$$\mathcal{L} := \mathbb{E}_{\mathbf{x}_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2]. \quad (3)$$

**Conditional Generation.** For conditional generation, a paired data  $(\mathbf{x}_0, \mathbf{y}_0)$  is given, and the aim is to model the conditional data distribution  $q(\mathbf{x}_0|\mathbf{y}_0)$ , where  $\mathbf{y}_0$  can be image class or text prompt. Conditional generation includes classifier guidance (Dhariwal and Nichol, 2021) and classifier-free guidance (Ho and Salimans, 2021). Classifier guidance requires training an extra classifier on noisy data at inference time to improve sample quality. For classifier-free guidance, no classifier needs to be trained. The denoising network  $\epsilon_\theta(\mathbf{x}_t|\mathbf{y}_0)$  simply conditions on the information encoded in  $\mathbf{y}_0$ . At inference time, with a guidance scale  $s$ , the modified score estimate is further in the direction of  $\epsilon_\theta(\mathbf{x}_t|\mathbf{y}_0)$  and away from the unconditional model  $\epsilon_\theta(\mathbf{x}_t|\emptyset)$  ( $\emptyset$  is a null token):

$$\hat{\epsilon}_\theta(\mathbf{x}_t|\mathbf{y}_0) = \epsilon_\theta(\mathbf{x}_t|\emptyset) + s \cdot (\epsilon_\theta(\mathbf{x}_t|\mathbf{y}_0) - \epsilon_\theta(\mathbf{x}_t|\emptyset)).$$

### 4 Proposed Model: EasyGen

We propose EasyGen, a model capable of processing multimodal inputs and generating multimodal outputs. It achieves easy multimodal generation by leveraging a bidirectional conditional diffusion model to effectively bridge the gap between different modalities and an LLM to comprehend multimodal tasks and produce textual responses containing cues for multimodal message creation. In the subsequent section, we outline the multimodal generation process of EasyGen.

#### 4.1 Pre-training BiDiffuser: A Bidirectional Conditional Diffusion Model

Since the text space of LLMs is discrete, to minimize the disparity between the output of a diffusion model and the input of LLMs, we leverage UniDiffuser, a unified diffusion model capable of transforming images into the discrete text space. During the training process, UniDiffuser injects noise  $\epsilon^x$  and  $\epsilon^y$  to a set of paired image-text data  $(\mathbf{x}_0, \mathbf{y}_0)$  and generates noisy data  $\mathbf{x}_{t^x}$  and  $\mathbf{y}_{t^y}$ , where  $0 \leq t^x, t^y \leq T$  represent two individual timesteps (perturbation levels). It then trains a joint denoising transformer U-ViT (Bao et al., 2023a)

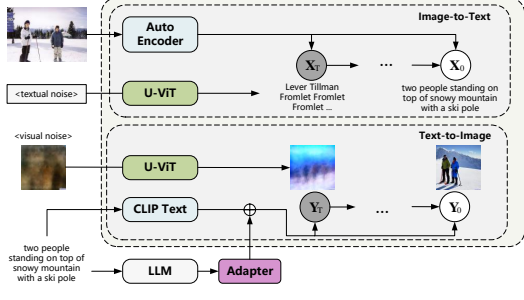


Figure 3: The training of BiDiffuser involves finetuning the denoising transformer U-ViT in UniDiffuser with a joint objective of image-to-text and text-to-image tasks.

$\epsilon_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}, t^x, t^y)$  to predict the noise  $\epsilon^x$  and  $\epsilon^y$  by minimizing the mean squared error loss:

$$\mathbb{E}_{\epsilon^x, \epsilon^y, \mathbf{x}_0, \mathbf{y}_0} [\|\epsilon^x, \epsilon^y - \epsilon_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}, t^x, t^y)\|^2],$$

where the output of  $\epsilon_\theta$  is the concatenation of the estimated noise  $\epsilon_\theta^x$  and  $\epsilon_\theta^y$ , i.e.,  $\epsilon_\theta = [\epsilon_\theta^x, \epsilon_\theta^y]$ .

By predicting  $\epsilon_\theta(\mathbf{x}_{t^x}, \mathbf{y}_{t^y}, t^x, t^y)$  for any  $t^x$  and  $t^y$ , UniDiffuser learns all distributions related to  $(\mathbf{x}_0, \mathbf{y}_0)$  simultaneously. This includes all conditional distributions:  $q(\mathbf{x}_0|\mathbf{y}_0)$  for text-to-image generation,  $q(\mathbf{y}_0|\mathbf{x}_0)$  for image-to-text generation, and those conditioned on noisy input, i.e.,  $q(\mathbf{x}_0|\mathbf{y}_{t^y})$  and  $q(\mathbf{y}_0|\mathbf{x}_{t^x})$ , for  $0 < t^x, t^y \leq T$ . Learning a conditional distribution  $q(\mathbf{x}_0|\mathbf{y}_{t^y})$  or  $q(\mathbf{y}_0|\mathbf{x}_{t^x})$  can be seen as learning a distinct task. From a multitask learning perspective, due to limited network capacity, learning many tasks simultaneously (i.e., fitting all distributions to a single network) may result in *task competition* or *task conflict*, ultimately leading to suboptimal performance in particular tasks such as  $q(\mathbf{x}_0|\mathbf{y}_0)$  and  $q(\mathbf{y}_0|\mathbf{x}_0)$ .

To resolve this issue and enhance the performance of both image-to-text and text-to-image generation tasks, we finetune UniDiffuser with exclusive emphasis on the two tasks:

$$\mathcal{L}_d = \mathbb{E}_{\epsilon^x, \epsilon^y, \mathbf{x}_0, \mathbf{y}_0} [\|\epsilon^x - \epsilon_\theta^x(\mathbf{x}_{t^x}, \mathbf{y}_0, t^x, 0)\|^2 + \alpha \|\epsilon^y - \epsilon_\theta^y(\mathbf{x}_0, \mathbf{y}_{t^y}, 0, t^y)\|^2],$$

where  $\alpha$  is a hyperparameter to balance the learning paces of the two tasks. As depicted in Figure 3, our training objective entails predicting the text  $\mathbf{y}_0$  based on the input image  $\mathbf{x}_0$  and vice versa, where the input conditions for the model are noise-free. We name the finetuned model “BiDiffuser”, signifying its specialized ability in bidirectional conditional generation.

## 4.2 Pre-training an Adapter to Enhance BiDiffuser’s SUR Capability

BiDiffuser uses the text encoder of CLIP, which is trained with image-text contrastive learning, limiting its semantic understanding and reasoning (SUR) ability for image generation. Drawing inspiration from Zhong et al. (2023), we utilize LLMs to enhance the SUR capability of BiDiffuser. Specifically, we design an adapter that employs the attention mechanism to integrate the semantic information from LLM’s last hidden state  $f_{\text{LLM}}(\cdot)$  into the CLIP text encoder  $f_{\text{CLIP}}(\cdot)$ . The adapter consists of a projection layer  $\text{MLP}(\cdot)$  and a cross-attention layer  $\text{Att}(\cdot)$ . Given a paired image-text data  $(\mathbf{x}_0, \mathbf{y}_0)$ , we can get  $y_{\text{sur}}$  with enhanced SUR via the adapter:

$$y_{\text{sur}} = \text{Att}(f_{\text{CLIP}}(\mathbf{y}_0)W^Q, \text{MLP}(f_{\text{LLM}}(\mathbf{y}_0))W^K, \text{MLP}(f_{\text{LLM}}(\mathbf{y}_0))W^V).$$

Then, the semantic input to BiDiffuser is the combination of  $y_{\text{sur}}$  and the CLIP text encoding of  $\mathbf{y}_0$ :

$$y_0 = \lambda \cdot y_{\text{sur}} + (1 - \lambda) \cdot f_{\text{CLIP}}(\mathbf{y}_0), \quad (4)$$

where  $\lambda$  is a balancing parameter. We train the adapter by freezing BiDiffuser and minimizing

$$\mathcal{L}_{\text{ada}} = \mathbb{E}_{\epsilon^y, \mathbf{x}_0} [\|\epsilon^x - \epsilon_\theta^x(\mathbf{x}_{t^x}, y_0, t^x)\|^2], \quad (5)$$

where  $\epsilon_\theta^x$  is not updated as BiDiffuser is frozen.

## 4.3 Image-to-Text Generation

BiDiffuser can convert images into vectors in the text space, facilitating alignment with the vector space of LLMs. In the following, we show how BiDiffuser can be integrated with LLMs to perform image-to-text generation tasks such as image captioning and visual question answering (VQA).

### 4.3.1 Aligning BiDiffuser with LLMs

We connect BiDiffuser and LLMs via a simple projection layer, which maps text embeddings obtained from the output of the diffusion model to the embedding space of LLMs. As shown in Figure 4, the alignment can take place either prior to the LLM (Pre-Align manner) or between its encoder and decoder components (Mid-Align manner).

**Pre-Align Manner.** As shown in Figure 4a, the projection layer is placed before the LLM to map the output of BiDiffuser (image representations) to the text embedding space of the LLM. The text embedding of the input image is then concatenated

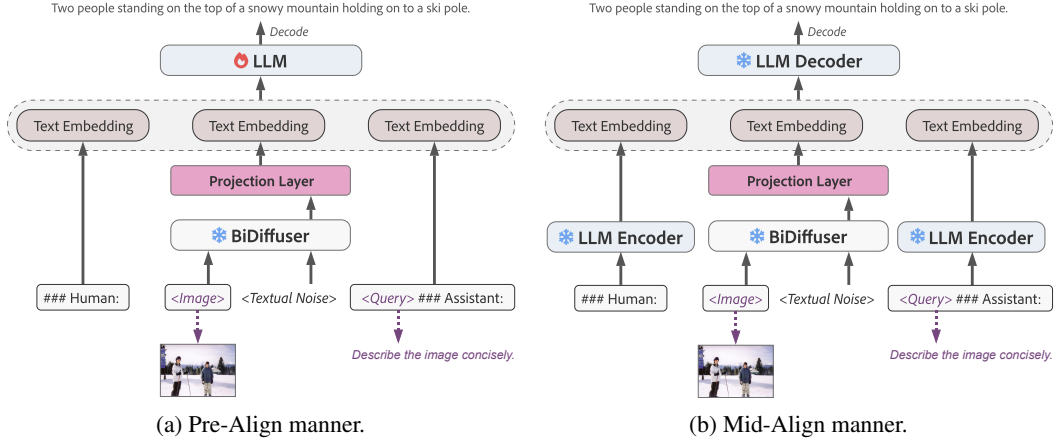


Figure 4: Two different ways of aligning BiDiffuser with LLMs.

with the embeddings of the textual instructions and fed to the LLM for decoding. To synchronize the text space of BiDiffuser with that of the LLM, we propose to use the image-grounded text generation (ITG) objective to drive the model to generate texts based on the input image by computing the auto-regressive loss:

$$\mathcal{L}_{\text{ITG}} = -\frac{1}{L} \sum_{l=1}^L \log p_{\phi}(w_l^g | w_{<l}^g, I, T_I), \quad (6)$$

where  $w^g = (w_1^g, \dots, w_L^g)$  represents the ground-truth caption of image  $I$  with length  $L$ ,  $T_I$  is the text instruction, and  $\phi$  denotes the model parameters, which include the parameters of the projection layer and the LLM.

**Mid-Align Manner.** As shown in Figure 4b, the projection layer is placed between the LLM’s encoder and decoder, aiming to map the output of BiDiffuser to the embedding space of the text that is encoded by the LLM’s encoder. Particularly, we argue that the output of BiDiffuser, once mapped by the projection layer and denoted as  $\mathbf{d}_{\text{diff}}$ , should align with the image caption that is encoded by the LLM’s encoder, denoted as  $\mathbf{d}_{\text{llm}}$ . Therefore, to accurately learn the alignment between the image and text representations, in addition to the ITG loss in Eq. 6, we also employ an image-text distance minimization (ITDM) loss:

$$\mathcal{L}_{\text{ITDM}} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{d}_{\text{diff}} - \mathbf{d}_{\text{llm}}\|_2^2, \quad (7)$$

$$\mathcal{L}_{\text{mid}} = \mathcal{L}_{\text{ITG}} + \mathcal{L}_{\text{ITM}}.$$

where  $N$  is the batch size, and  $\mathcal{L}_{\text{mid}}$  is the overall loss. In this manner, the model parameters  $\theta$  only include the parameters of the projection layer.

After the alignment, EasyGen gains the capability of zero-shot image-to-text generation, including tasks such as image captioning and VQA.

#### 4.3.2 Instruction-Tuning LLMs

When aligning BiDiffuser with an LLM, we perform instruction-tuning on the LLM to equip it with the capability of understanding multimodal tasks. We designed different instructions for different LLMs, as shown in Table 12. General instruction template is denoted as follows:

USER: `<Img><image></Img>` + Instruction. Assistant: `<answer>`.

For the `<image>` placeholder, we substitute it with the output of BiDiffuser. To avoid over fitting to the specific task and counter the model’s inclination to generate excessively short outputs, we have devised specific instructions (see Table 11), which enable the LLM to produce concise responses when necessary. For different tasks, the distinct instruction templates are as outlined in Appendix F.

#### 4.4 Text-to-Image Response Generation

Most of existing multimodal models, including the BLIP series and LLaVA series are unable to provide a multimodal response as they are primarily designed to generate only textual outputs. On the other hand, Emu (Sun et al., 2023) takes a unified approach to predict the subsequent visual or textual token in an auto-regressive manner, but it is heavily reliant on vast quantities of training data. Contrary to the limitations of these existing models, EasyGen, by leveraging the bidirectional generation capability of BiDiffuser and the inference capability of LLMs, can produce accurate and high-quality visual response with ease.

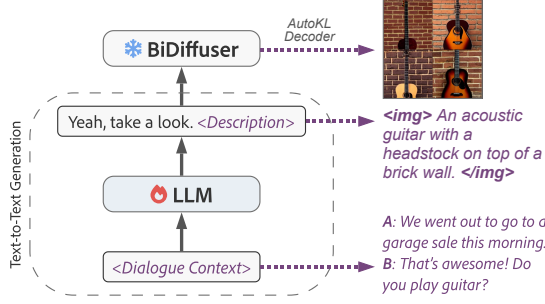


Figure 5: Text-to-image generation by EasyGen. LLM generates the response and description of the image. BiDiffuser generates images based on the description.

To tackle multimodal response generation tasks such as PhotoChat (Zang et al., 2021), we first leverage the MLLM to generate detailed image captions based on dialogue context. Then, we employ BiDiffuser to create the corresponding images with the produced captions. Specifically, we replace the image featured in the dialogue with its corresponding descriptive caption, encapsulating it with task-specific tokens `<Img>`, `</Img>` and constructing the following instruction templates:

USER: Dialog history. Assistant: `<response>` + `<Img><caption></Img>`.

When `<caption>` appears in response, it represents the generated description of the image. So we can use LLM’s original auto-regressive training objective. Specifically, we compute the probability of the target caption by:

$$\mathcal{L}_{t2t} = -\frac{1}{L} \sum_{l=1}^L \log p_{\theta}(w_l^c | w_{<l}^c, H), \quad (8)$$

where  $w^c = (w_1^c, \dots, w_L^c)$  represents the caption of image  $x_0$  with length  $L$ ,  $H$  is the dialogue history, and  $\theta$  denotes the LLM’s parameters. Considering the potential for alignment deviation in discrete text alone, given the description of the image  $y_0$ , we utilize  $y_0$ , which is the combination of the SUR adapter’s output and the CLIP text encoder’s output, as the conditional component of the diffusion model. This directly contributes to the denoising process. The loss function for the denoising process of a noisy image  $x_{tx}$  is formulated in a way that is similar to Eq. 5:

$$\mathcal{L}_{t2i} = \mathbb{E}_{\epsilon^y, x_0} [\|\epsilon^x - \epsilon_{\theta}^x(\mathbf{x}_{tx}, y_0, t^x)\|^2], \quad (9)$$

where  $\epsilon_{\theta}^x$  is not updated and we only train the parameters of LLM and adapter. The overall loss for text-to-image task is:

$$\mathcal{L}_{all} = \mathcal{L}_{t2i} + \mathcal{L}_{t2t}. \quad (10)$$

Training with the instruction data enables our model to not only produce text responses but also perform image intent classification and generate image captions that BiDiffuser can interpret.

## 5 Experiments

### 5.1 Experimental Setup

We initialize encoder-decoder LLM from FlanT5-XL or decoder-only LLM from Vicuna-7B, along with the utilization of the diffusion module from BiDiffuser. During the alignment process, we maintain the frozen state of the BiDiffuser. The statistics of the datasets for pre-training, alignment and instruction-tuning can be found in Appendix B. For the image captioning task, EasyGen is evaluated on both the MS-COCO (Lin et al., 2014) Karpathy test set and the NoCaps (Agrawal et al., 2019) validation set. For the VQA task, we evaluated on OK-VQA (Marino et al., 2019) validation set and GQA (Hudson and Manning, 2019) test-dev set.

To adapt the model for multimodal dialogue generation, we fine-tune the LLM and projection layers on the PhotoChat dataset. We incorporate photo-sharing activities into the dialogue context by generating `<Img><caption></Img>`, and utilize cross-entropy loss exclusively for fine-tuning the multimodal generation task. Given the limited expressiveness of image descriptions in the PhotoChat dataset, as evidenced by Figure 7 in Appendix I, we regenerate image annotations in a text format similar to that used in MS-COCO.

### 5.2 Evaluation

We evaluate EasyGen on various vision-language tasks including image captioning (MS-COCO (Lin et al., 2014), NoCaps (Agrawal et al., 2019)), visual question answering (OK-VQA (Marino et al., 2019), GQA (Hudson and Manning, 2019)), and multimodal dialog generation (PhotoChat (Zang et al., 2021)). We use BLIP (Li et al., 2022), Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), InstructBlip (Dai et al., 2023), MiniGPT-4 (Zhu et al., 2023), and LLaVA (Liu et al., 2023b) as baselines for image-to-text tasks, and Maria (Liang et al., 2021) and Divter (Sun et al., 2021) as baselines for the multimodal response generation task. See details in Appendix C and E.

### 5.3 Overall Results

Tab. 1 presents the evaluation results for each baseline and our models on MS-COCO and VQA (zero-

Model	Dataset Size		NoCaps (val)		COCO (Karpathy)		OK-VQA	GQA
	PT	IT	CIDEr	SPICE	BLEU@4	CIDEr	Accuracy	Accuracy
BLIP (Li et al., 2022)	129M	-	113.2	14.8	40.4	136.7	-	-
Flamingo (Alayrac et al., 2022)	1.8B	-	-	-	-	138.1	50.6	-
BLIP-2 OPT-6.7B (Li et al., 2023b)	129M	-	121.0	15.3	<b>43.5</b>	145.2	36.4	36.4
BLIP-2 FlanT5XL (Li et al., 2023b)	129M	-	121.6	<b>15.8</b>	42.4	144.5	39.4	44.4
InstructBlip 7B (Dai et al., 2023)	129M	1.2M	<b>123.1</b>	-	40.8	140.7	61.0*	49.2*
MiniGPT-4 (Zhu et al., 2023)	-	5M	42.4	-	-	-	37.5	30.8
LLaVA (Liu et al., 2023b)	558K	158K	33.1	-	7.9	30.0	54.4	41.3
<b>EasyGen FlanT5XL</b>	169K	90K	121.2	15.5	<b>43.5</b>	<b>145.7</b>	41.1	37.2
<b>EasyGen Vicuna-7B</b>	169K	90K	121.8	<b>15.8</b>	42.4	144.6	45.2	<b>44.6</b>

Table 1: Evaluations of EasyGen and baselines on various **image understanding** tasks. PT, IT indicate sample sizes in the pretraining and instruction tuning stages respectively. EasyGen’s results on NoCaps, OK-VQA and GQA were obtained in a zero-shot setting. \* denotes that the model was trained on other VQA datasets.

Model	Response Generation			Image
	BLEU-1/2	PPL↓	ROUGE-L	FID↓
Divter Sun et al.	6.5/1.7	59.6	5.69	29.16
Maria Liang et al.	13.8/9.2	48.7	15.17	-
<b>EasyGen FlanT5</b>	22.3/18.7	13.3	17.24	10.30
<b>EasyGen Vicuna</b>	<b>23.6/19.9</b>	<b>11.3</b>	<b>18.85</b>	<b>9.72</b>
+ w/o adapter	-	-	-	10.16

Table 2: Evaluation on the **PhotoChat** dataset.

MLLM	Sample Size	Cosine Similarity ↑	MSE ↓
MiniGPT-4	5M	0.0016	6.2031
LLaVA v1.5	558K	-0.0026	0.8433
Emu	2B	0.0054	0.4062
EasyGen	169K	0.0128	0.0338

Table 3: **Data efficiency.** Avg. Cosine similarity and mean square error between the projected representations and their respective captions embedded by LLM.

shot) datasets. EasyGen outperforms most of the baseline models on both the COCO test set and NoCaps validation set (zero-shot transfer). Despite being pre-trained on a small dataset (MS-COCO), EasyGen’s performance on the image captioning generation task is comparable to models trained on larger datasets. Additionally, on the OK-VQA and GQA datasets, EasyGen demonstrates improved performance compared to other models of a similar scale, achieving higher accuracy even with a simple greedy search decoding method.

In Tab. 2, the evaluation results on the PhotoChat dataset are presented. Our method exhibits clear advantages in terms of PPL, indicating strong performance on response generation task. Because of the image descriptions in the PhotoChat dataset are overly concise, we utilized EasyGen to regenerate the image descriptions, which improved the performance of our model on image generation compared to other models. Additionally, with the adapter, EasyGen is capable of generating images of superior quality.

## 5.4 Ablation Study

In Tab. 4, we examine the impact of freezing/tuning BiDiffuser and the LLM. It can be observed that frozen Mid-Align method outperforms Pre-Align method in image captioning, which shows ITDM loss function is effective. However, the frozen Mid-Align method exhibits inferior performance in the VQA task. We hypothesize that this is due to the integration of mid-aligned target image features with query information, and the projection layer is insensitive to instruction information. We conduct instruction-tuning on Pre-Align T5 and Vicuna. Compared to models at the same scale, these instruction-tuned models achieve superior results.

## 5.5 Data Efficiency in Training

In Tab. 3, we examine the data efficiency of different image encoders for alignment with LLMs. EasyGen uses BiDiffuser, which maps images to the text space, simplifying alignment with LLMs. To assess the quality of visual representations, we measured the distance between the projected representations and their respective captions embedded by an LLM. We randomly selected 1,000 images with their corresponding captions from the MSCOCO dataset. The results show that our model, EasyGen, aligns significantly better with the LLM compared to other CLIP-based MLLMs, despite using less data for alignment. This indicates the effectiveness of our approach in achieving strong alignment with LLMs.

## 5.6 Image Generation Quality

Tab. 6 evaluates the generated image’s quality of MLLMs on MS-COCO validation set, using 30K randomly selected prompts to compute the FID score on generated images. To confirm the efficacy of our approach, we fine-tuned our method on a

LLM	Diffusion Model	Alignment	NoCaps	COCO(Karpathy)			OK-VQA
			CIDEr	SPICE	BLEU@4	CIDEr	Accuracy
❄️ T5	UniDiffuser	Pre-Align	62.4	18.0	26.8	90.7	33.0
🔥 T5	BiDiffuser	Pre-Align	119.1	<b>25.5</b>	42.6	145.1	<b>41.1</b>
❄️ T5	BiDiffuser	Mid-Align	121.2	25.1	43.5	<b>145.7</b>	31.5
🔥 T5	BiDiffuser	Mid-Align	121.5	25.3	<b>43.6</b>	<b>145.7</b>	36.4
🔥 Vicuna-7B	BiDiffuser	Pre-Align	<b>121.8</b>	24.9	42.4	144.6	<b>45.2</b>
❄️ Vicuna-7B	BiDiffuser	Pre-Align	119.0	24.6	40.3	140.3	42.7

Table 4: **Ablation study** on image captioning and VQA tasks. 🔥 / ❄️ denotes tuning/freezing the LLM.

Model	IT	VQAv2 (test-dev)	TextVQA	MMBench (test)
MiniGPT-4 (Zhu et al., 2023)	5M	-	19.4	23.0
InstructBLIP Vicuna-7B (Dai et al., 2023)	1.2M	-	50.1	33.9
LLaVA-1.5 Vicuna-7B (Liu et al., 2023a)	665K	78.5	58.2	65.2
LLaVA-1.5 Vicuna-13B (Liu et al., 2023a)	665K	80.0	61.3	67.8
EasyGen Vicuna-7B w/ ViT-L	251K	79.4	57.9	63.9
LLaVA-1.5 Vicuna-7B w/ EasyGen	665K	80.2	58.8	66.1
LLaVA-1.5 Vicuna-13B w/ EasyGen	665K	<b>80.5</b>	<b>61.5</b>	<b>69.2</b>

Table 5: Evaluation of **EasyGen variants** and baselines on more complex VQA tasks and the latest MMBench benchmark. “w/ EasyGen” means incorporating the core components of our model into existing models as depicted in Figure 6 in Appendix F. EasyGen variants rank among the top models on the leaderboard of MMBench.

MM-Model	FID ↓	Diffusion Model	FID ↓
<i>Zero-Shot</i>			
NExT-GPT	11.28 (+0.07)	SD	11.21
Emu	11.66 (+1.73)	SD v1.5	9.93
EasyGen	9.16 (-0.55)	UniDiffuser	9.71
+ w/o adapter	9.52 (-0.19)	UniDiffuser	9.71
<i>Fine-tuned on MS-COCO</i>			
EasyGen	7.68 (-0.44)	UniDiffuser	8.12
+ w/o adapter	7.89 (-0.23)	UniDiffuser	8.12

Table 6: Comparing the **image generation quality** of end-to-end MLLMs and their corresponding diffusion models on the MS-COCO validation set ( $256 \times 256$ ). Our EasyGen surpasses the original diffusion model, while other MLLMs fall short in comparison.

portion of the original data (LIAON-COCO) and the MS-COCO train set, respectively. While other models resulted in a decrease in image generation performance compared to the corresponding diffusion model, EasyGen outperformed the original UniDiffuser due to the fine-tuned BiDiffuser and the adapter module. Furthermore, Tab. 7 provides CLIP-T scores from ImagenHub. We notice similar trends to the results in Tab. 6 using the FID indicator. This suggests that our method can better align LLM with diffusion model’s text space.

## 5.7 Extendability

Tab. 5 explores the extensibility of our method from two aspects. Firstly, we aim to enhance the performance of EasyGen on complex tasks such as VQA and OCR by integrating more powerful visual encoders. Considering the potential information dilu-

MM-Model	CLIP-T ↑	Diffusion Model	CLIP-T ↑
NExT-GPT	0.259 (-0.031)	SD	0.290
Emu	0.262 (-0.023)	SD v1.5	0.285
Emu2	0.266 (-0.023)	SD XL	0.289
EasyGen	9.16 (-0.55)	UniDiffuser	9.71

Table 7: Comparing the **CLIP-T score** of end-to-end MLLMs and their corresponding diffusion models on the ImagenHub.

tion or omission when using BiDiffuser to convert images to text space, we choose to integrate CLIP ViT-L/14 as the image encoder (as depicted in Figure 6 in the Appendix). During this process, we freeze CLIP and BiDiffuser while fine-tuning the parameters of the LLM and projection layers. The results presented in Tab. 5 include performance on traditional short QA and the modern benchmark MMBench (Liu et al., 2023c). With CLIP ViT-L, EasyGen’s performance is better than LLaVA on the VQAv2 dataset, demonstrating that BiDiffuser can effectively assist LLM in understanding images. Secondly, we investigate the plug-and-play capability of BiDiffuser, as it can also be integrated into other MLLMs (with the same LLMs) to improve their performance. The results demonstrate that with BiDiffuser, LLaVA-1.5 could achieve better performance. We speculate that BiDiffuser provides guidance information to MLLMs, enabling them to better understand the details of CLIP encoded images.

## 6 Conclusion and Future Work

We have introduced EasyGen, a model that facilitates multimodal understanding and generation. Compared to existing models, EasyGen offers a more efficient solution by employing BiDiffuser, a bidirectional diffusion model. This allows for more effective modal interactions, handling both image-to-text and text-to-image generations by the fusion of BiDiffuser and LLMs. Additionally, EasyGen can be easily integrated into existing advanced multimodal LLMs to further boost their performance. In the future, we will explore adapting EasyGen to perform a broader range of multimodal tasks, including subject-driven image generation, image editing, and controlled generation.

## 7 Limitations

This section aims to highlight the limitations of our work and provide further insights into the research in this area. Our model relies on diffusion for multimodal interaction, which means that the text-to-image and image-to-text processes may take longer. In our experiments, we tested the performance of our model on one A100 (80G) GPU. During inference, using 1000 image-caption pairs, EasyGen took approximately 2.95 seconds for the caption generation task (with the diffusion module taking about 2.41 seconds) and around 4.96 seconds to generate an image. We believe it would be beneficial to explore more efficient sampling methods, such as DPM-Solver++ (Lu et al., 2022), to improve the overall efficiency of EasyGen. Furthermore, EasyGen may not be seamlessly adaptable for jointly fine-tuning the BiDiffuser and Language Model without altering the BiDiffuser’s sampling mechanism. But based on our experimental findings, that joint fine-tuning of the BiDiffuser and LLM is not necessary for handling complex tasks. This observation aligns with many established approaches that do not require fine-tuning of their corresponding image encoders.

## Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was partially supported by the grant of HK ITF ITS/359/21FP.

## References

- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. No-caps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.
- Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023a. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22669–22679.
- Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. 2023b. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*.
- Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- 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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

- Runpei Dong, Chunrui Han, Yang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Manat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. 2022. From images to textual prompts: Zero-shot vqa with frozen large language models. *arXiv preprint arXiv:2212.10846*.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023a. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*.
- Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.
- Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A visual experience powered conversational agent. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 5596–5611.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference*

- on computer vision and pattern recognition, pages 3195–3204.
- Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR.
- Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2023. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugging-gpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.
- Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2021. Multi-modal dialogue response generation. *arXiv preprint arXiv:2110.08515*.
- Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.
- Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.
- Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit

- Som, et al. 2022b. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multi-modal llm. *arXiv preprint arXiv:2309.05519*.
- Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. 2023. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023b. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*.
- Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. 2023. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 567–578.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## A Ethics Statement

We adhere to the ACL Ethics Policy and have conducted our research using publicly available repositories and datasets. Our primary focus is on investigating the integration of diffusion models and LLMs for multimodal generation. Therefore, the results should be seen as AI-generated content. While we have not observed deliberate harmful content, the model has the potential to generate such content if triggered. We have taken steps to minimize this risk through fine-tuning on public datasets, but caution is still exercised. In future, we will prioritize improving downstream performance and exploring methods to enhance control over the generation process. To ensure reproducibility and support future research, we have made all resources publicly available and provided proper citations to previous research within the code.

## B Datasets

We test the effectiveness of EasyGen by experimenting on different tasks including image captioning, visual question answering (VQA), and multimodal dialogue tasks. Table 8 shows the statistics of the pre-training datasets for BiDiffuser, alignment and VQA tasks.

We use the MS-COCO (Lin et al., 2014) dataset for image captioning. Following BLIP-2 (Li et al., 2023b), we fine-tune EasyGen on MS-COCO and evaluate its performance on the Karpathy test set and the NoCaps (Agrawal et al., 2019) validation set. In MS-COCO, each image typically has five captions that convey similar meanings. The training set consists of 82,783 images with 414,113 captions, while the COCO Karpathy test set has 5,000 images and the NoCaps validation set has 4,500 images.

For multimodal dialogue, we utilize the PhotoChat (Zang et al., 2021) dataset, which is a high-quality dataset consisting of 10,917 images and 12,286 dialogues. Each dialogue is associated with a user image and its corresponding text description. The dataset is divided into 10,286 training instances, 1,000 development instances, and 1,000 testing instances. Moreover, PhotoChat includes photo-sharing activities, defined as the process of creating `<Img><caption></Img>` in this study. Each conversation in PhotoChat is broken down and constructed into multiple samples so that each round of responses can be learned. Specifically, we regard the first three turns as the dialog context,

and the subsequent turns as the prediction targets. By converting the dialogues of this dataset into the form mentioned in 4.4, we obtained 49,240 train, 4,792 dev, and 4,836 test dialogue pairs.

For the VQA task, we conduct a quantitative evaluation on both the OK-VQA (Marino et al., 2019) validation set (5,046 questions) and the GQA (Hudson and Manning, 2019) test-dev set (12,578 questions). As shown in Table 4, for the frozen LLM, following BLIP-2, we employ the length penalty in beam search to encourage short answer generation. On the contrary, for the tuned LLM, we use the VQA instructions (as shown in Table 10) to do instruction tuning during the alignment process. The data for instruction tuning is constructed by randomly selecting 5K data from VQAv2 (Goyal et al., 2017) and 5K data from Visual Dialog (Murahari et al., 2019) training set.

## C Baselines

We compare our proposed model with the following state-of-the-art baselines:

**BLIP** (Li et al., 2022) is a multimodal mixture of encoder-decoder. It can be used as an image-based text encoder or decoder. We use it to perform caption generation and VQA tasks.

**BLIP-2** (Li et al., 2023b) is pre-trained through bootstrapped learning from frozen visual encoder and LLMs using an efficient pre-training strategy.

**Flamingo** (Alayrac et al., 2022) incorporates new cross-attention layers into Chinchilla language model (Hoffmann et al., 2022) to inject visual features, and pre-trains the new layers on billions of image-text pairs. We use it to perform caption generation and VQA tasks.

**InstructBlip** (Dai et al., 2023) is a vision-language instruction tuning framework that is trained with BLIP-2 and capable of solving various visual language tasks.

**MiniGPT-4** (Zhu et al., 2023) utilizes a single projection layer to align visual information from a pre-trained vision encoder with an LLM. It employed the same visual encoder as used in BLIP-2.

**LLaVA** (Liu et al., 2023b) employs a solitary projection layer to convert image features extracted from the pre-trained CLIP-ViT-L/14 visual encoder into the language embedding space of Vicuna.

**Maria** (Liang et al., 2021) is a neural conversation agent which can leverage visual world experiences sourced from a vast image index. It possesses the ability to fetch a relevant image specific to the con-

Data types	Dataset	Size	BiDiffuser	Alignment	Fine-tuning
Caption	MS-COCO caption (Lin et al., 2014)	83K	✓	✓	✗
	Visual Genome (Krishna et al., 2017)	86K	✓	✗	✗
Multimodal instruction	LLaVA dataset (Liu et al., 2023b)	80K	✗	✓	✓
VQA	VQAv2 (Goyal et al., 2017)	83K	✗	-	✓
	AOK-VQA (Schwenk et al., 2022)	66K	✗	✗	✓
OCR-related tasks	Text Captions (Sidorov et al., 2020)	22K	✗	✗	✓
	TextVQA (Singh et al., 2019)		✗	✗	✓

Table 8: Description of datasets used in our alignment and VQA fine-tuning stages. Noting that in alignment process, we used 5K images from VQAv2 dataset.

	Dataset	Task	Split	Metric
<b>Image-to-Text</b>	MS-COCO (Lin et al., 2014)	Image captioning	Test	CIDEr, BLEU, SPICE
	NoCaps (Agrawal et al., 2019)	Image captioning	Val	CIDEr, SPICE
	OK-VQA (Marino et al., 2019)	VQA	Val	Accuracy
	GQA (Hudson and Manning, 2019)	VQA	Test	Accuracy
<b>Multimodal Generation</b>	PhotoChat Zang et al., 2021	Image dialogue	Test	PPL, BLEU, ROUGE, FID

Table 9: Summary of the evaluation datasets and metrics.

version and extract visual knowledge from it.

**Divter** (Sun et al., 2021) focuses on exploring multimodal dialogue generative models. Given the dialogue context, this model first generates a text response or image description and then generates an image according to the description.

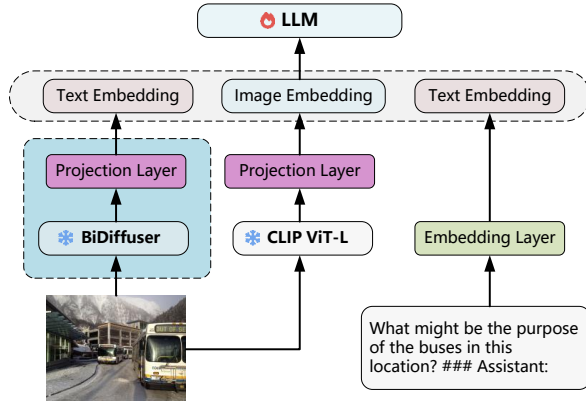


Figure 6: Model’s architecture for VQA finetuning. The module with blue background is referred to as BiDiffuser, while the rest is the architecture of MLLM using CLIP as the image encoder (such as LLaVA).

## D Evaluation

For evaluating the quality of text generation, we utilize metrics such as BLEU, Rouge-L, Accuracy, and PPL (Perplexity). Additionally, following the approach of Vicuna (Chiang et al., 2023) and LLaVA (Liu et al., 2023b), we employ ChatGPT to assess the generated responses from our model. Specifically, for the image captioning task,

we randomly select 30 images from the MS-COCO Karpathy split and then let ChatGPT score the responses generated by EasyGen and the baseline models. ChatGPT evaluates the models’ responses based on relevance, details, and accuracy and assigns an overall score between 1 and 10, with a higher score indicating better performance. To evaluate the quality of image generation, we use the Frechet Inception Distance (FID) score (Heusel et al., 2017), which measures the divergence between two multivariate normal distributions.

## E Implementation Details

**LLM** During the alignment process, we utilize the AdamW optimizer with  $\beta_0 = 0.9$ ,  $\beta_1 = 0.99$ , and weight decay of 0. The LLMs are trained with a cosine learning rate of  $2e-5$  and a warmup rate of 0.03. We use a batch size of 96 for the frozen LLMs and 32 for the tuned LLMs. During training, we convert the LLMs (FlanT5XL/Vicuna-7B) to BFloat16/FP16 and BiDiffuser to FP16. During the VQA tuning process, we use CLIP ViT-L/14 336px as additional image encoder. We finetune EasyGen on mixture datasets for 1 epoch with a batch size of 32. We adopt the AdamW optimizer with  $\beta = (0.9, 0.99)$  with the learning rate is  $2e-5$ . We use a cosine learning rate decay with a learning rate is  $2e-5$  and warmup ration is 0.03.

**Diffusion Module** We inherit the settings from UniDiffuser and utilize pre-trained weights from its checkpoint for our text-to-image generator. The model is fine-tuned on the MS-COCO and VG

Task	Instruction Template
Image Captioning	USER: <image>+random[query] Assistant:
LLaVA 80K	USER: Please answer question from this image: <image> Question: <question> Assistant: USER: Image: <image> Question: <question> Assistant: USER: Answer question <question> through the image <image> Assistant:
Multimodal Dialogue	USER: Dialog history+<photo>+Dialogue history Assistant:
VQA	USER: Image: <image> Question: <question> Short answer: Assistant: USER: Image: <image> Question: <question> Answer the option's letter. Assistant:

Table 10: Examples of task instruction templates. <image> represents the input image, <question> denotes the question in the VQA and LLaVA 80K dataset, and <photo> is the image description of the input image.

dataset, which contains images with a resolution of  $512 \times 512$ , for 10 epochs with a batch size of 312. For all of our sampling processes, we employ DPM-Solver with 50 steps.

1. Describe the image **concisely**.
2. Provide a **brief** description of the given image.
3. Can you describe this image **briefly**?
4. Provide a **summary** of visual elements depicted in the image.
5. Give me the essential characteristics of the photograph in a **concise** manner.
6. Rephrase the image depicted in a **concise** manner.
7. Describe the objects in this image **no in detail**.
8. Please introduce the image for me **briefly**.
9. Give me the image's short descriptions.
10. Please provide a **general** depiction of the image presented.

Table 11: For the image captioning task, a query instruction is randomly selected.

## F Instruction Tuning

We list the instructions for different tasks in the main paper in Table 10. Specifically, the queries used to describe image contents are presented in Table 11. Table 10 shows the templates used in Vicuna, if the LLM is FlanT5, kindly use “Human” to substitute “USER” in the instruction templates. Model architecture for VQA finetuning is shown in Figure 6. EasyGen integrates the outputs of BiDiffuser with images encoded by CLIP ViT-L/14. We freeze CLIP and BiDiffuser while only tuning the parameters of the LLM and projection layers.

## G Training Efficiency

Table 13 summarizes the key factors in training EasyGen. The training process of EasyGen is computationally efficient, especially with the utilization of the parameter-efficient fine-tuning method LoRa (Hu et al., 2021). To enable multimodal response generation, we further train the aligned EasyGen. This process entails fine-tuning the LLM (FlanT5XL) on the PhotoChat dataset for 2 epochs, which typically requires approximately 4 A100 (80G) GPU hours.

### Different Instruction Templates:


	
Caption Generation Response	<b>Provide a brief description of the given image. Assistant:</b> Buses parked with a snow mountain view behind them.
VQA Response	What numbers are displayed on the front of the bus on the right? <b>Short answer: Assistant:</b> 6044
Multimodal Dialogue Response	What might be the purpose of the buses in this location? <b>Assistant:</b> It is plausible that they are in this location for multiple reasons. Some possible reasons might be: 1)...)2)...

Table 12: Examples of different instructions with different output formats. We use bold fonts to indicate different instructions.

## H Impact of Alignment Manners

In Table 14, we investigate the impact of different alignment manners on EasyGen. After removing the ITDM loss, the performance is slightly weaker than the original model. It is evident that the MSE Loss can help to align the semantic spaces of the two models. Furthermore, the performance of the model will drop significantly after removing the cross-entropy loss, suggesting that constraints via the language model play a key role.

## I More Qualitative Results

We present several instances on PhotoChat dataset in Figure 7 and the image-captioning task in Figure 8. In Figure 9, 11, 10, we compare EasyGen with state-of-the-art multimodal language models. The responses of MiniGPT-4, LLaVA, mPLUG-owl and InstructBlip are obtained from their official demos. Moreover, in Figure 12, 13, we show EasyGen’s ability to accept multimodal inputs and generate multimodal responses.

Model	Trainable Param.	Training Images	Training Cost
<i>Pre-training</i> BiDiffuser	952M	169K	120 (A100 80GB) GPU hours
<i>Alignment</i>			
Projection Layers + 🌸 T5XL	4M	163K	20 (RTX3090 24GB) GPU hours
Projection Layers + 🍷 T5XL	3B	173K	20 (A100 80GB) GPU hours
Projection Layers + 🍷 Vicuna 7B	7B	173K	72 (A100 80GB) GPU hours
Projection Layers + 🍷 Vicuna 7B(LoRa)	610M	173K	20 (A100 80GB) GPU hours

Table 13: EasyGen’s trainable parameters, training data size, and training cost during alignment process.

Model	NoCaps (val)		COCO (Karpathy)			OK-VQA	GQA
	CIDEr	SPICE	SPICE	BLEU@4	CIDEr	Accuracy	Accuracy
<b>EasyGen Mid-Align FlanT5XL</b>	121.2	<b>15.5</b>	<b>25.1</b>	<b>43.5</b>	<b>145.7</b>	31.5	22.6
+ w/o ITDM	118.6	15.3	24.8	42.2	141.5	-	-
+ w/o ITG	93.2	12.9	23.0	35.1	127.6	-	-

Table 14: Ablation studies on the instruction-tuning process and loss functions.

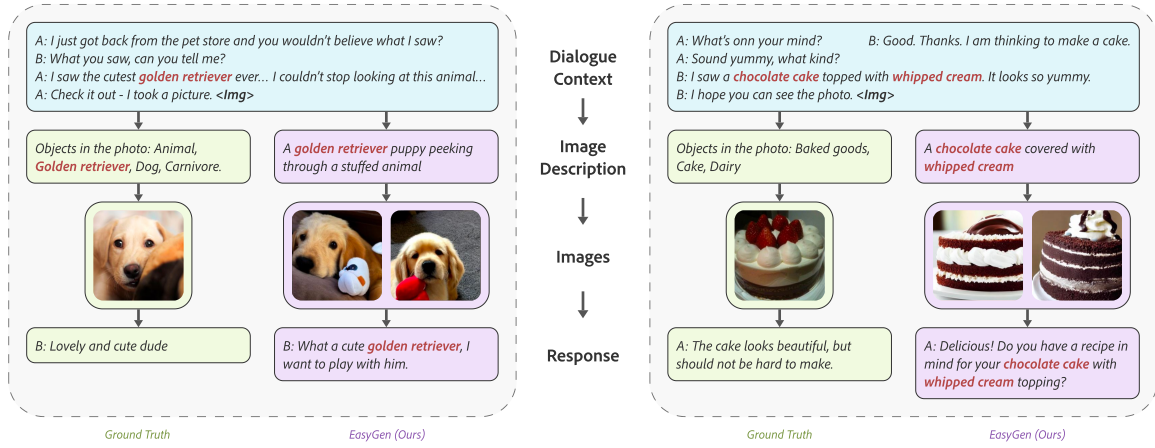


Figure 7: Examples of the generated responses on PhotoChat dataset. The text highlighted in red indicates the objects present in the image. The turns prefixed with A/B denote the given context.



EasyGen: A giraffe eating grass on a green grass field.



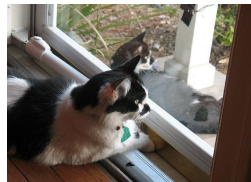
EasyGen: A kitchen with a stove and oven in a fireplace.



EasyGen: Two children are riding on skis with their parents in a snowy street.



EasyGen: A snowboarder flying through the air on a frozen mountain slope.



EasyGen: Two cats separated by the window are looking forward at the same time.



EasyGen: A long wooden bench with pictures of the world.

Figure 8: Examples of image captioning results by EasyGen.



Please tell me the information about this picture.

EasyGen (ours):

This painting is a classic representation of a young girl from the 17th century. The girl is depicted wearing a blue and white dress, with a pearl earring in her ear. The painting is a masterpiece of Dutch Golden Age art, a period of great cultural and economic prosperity in the Netherlands. The girl in the painting is likely depicted as a middle-class citizen, and the painting may have been commissioned as a portrait or for a specific purpose, such as a family crest or a religious painting. The painting is known for its attention to detail, vivid colors, and beautifully rendered textures. It is a true masterpiece of Dutch art, and a must-see for anyone interested in the history of art.

mPLUG-Owl:

The image is a painting depicting a young girl, likely a Dutch girl in a yellow dress, with her hair tied up in braids and a gold earring. She is looking directly into the viewer's eyes, creating an intimate connection with the observer. The painting is set in an indoor scene with a table and chairs visible in the background. The girl is positioned in front of the table, and there is an additional person standing to her left. There are also several other people in various positions around the room, suggesting a social gathering or a family setting.

LLoVA:

The image is a portrait of a woman wearing a blue headscarf, a yellow dress, and a pearl earring. She is looking to her left, and her outfit and accessories create a visually appealing contrast against the background. The woman's attire and appearance suggest a sense of style and elegance.

MiniGPT-4:

The painting depicts a young woman with a pearl earring looking out into the distance. The woman is wearing a yellow dress with a blue and white headband. The background is a mix of blue and green, with a few trees visible in the distance. The overall mood of the painting is peaceful and contemplative. The painting is likely a representation of a woman reflecting on her thoughts or feelings.

Generating some pictures like this.

EasyGen (ours):



mPLUG-Owl:

Here are a few examples of paintings that are similar in style to the one you described: 1. "Girl with a Pearl Earring" by Johannes Vermeer; 2. "The Procuress" by Pieter de Hooch; 3. "The Artist's Mother" by Gerard van Honthorst .... These paintings are all from the 17th century and depict women or groups of people in various settings, often with a focus on everyday life. They are known for their attention to detail and realistic depiction of light and shadow.

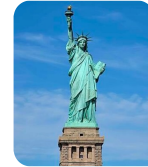
LLoVA:

I'm sorry, as an AI language model, I am not able to generate images. However, I can suggest some ideas for paintings that share a similar style to the portrait of the woman in the blue headscarf, yellow dress, and pearl earring by Johannes Vermeer:

MiniGPT-4:

Here are some pictures that resemble the painting, featuring young girls with their hair up in braids and wearing gold earrings: A young girl in a yellow dress and braided hair, looking directly into the viewer's eyes. Another young woman with her hair in similar braid styles, wears a white dress with gold accents, and has a gold necklace. Two girls in matching yellow outfits, with braiding and gold jewelry, standing together and looking into each other's eyes, creating a warm and inviting atmosphere.

Figure 9: In this case study, for the first question, EasyGen can give an accurate answer including the background information of the image. With the image generation ability of BiDiffuser, EasyGen can generate visual responses.



Where can I see this object appeared in this image?

EasyGen (ours):

The Statue of Liberty is located in New York City, United States. It is a famous landmark and tourist attraction, and you can see it in person by visiting Liberty Island in New York Harbor.

mPLUG-Owl:

You can see this blue and white 3D rendering of the Statue of Liberty located in the United States. The image features an artistically detailed representation of the famous landmark, which stands tall and proud in the skyline. The statue represents the ideals of freedom, democracy, and human rights. The scene is captured in a blue sky, emphasizing the beauty and grandeur of the iconic structure.

LLaVA:

You can see the Statue of Liberty in the image, standing tall and proud in front of a blue sky.

I plan to travel here, what should I pay attention to?

EasyGen (ours):

When visiting the Statue of Liberty, there are a few things you should keep in mind:

1. The Statue of Liberty is a colossal neoclassical sculpture on Liberty Island in New York Harbor within New York City, United States. It was a gift from the people of France to the people of the United States and was dedicated on October 28, 1886, to commemorate the centennial of the signing of the United States Declaration of Independence.
2. The statue is 301 feet (92 meters) tall, including its pedestal and base, and is made of iron and steel. It was designed by the French sculptor Frederic Auguste Bartholdi and the American architect Emmanuel Louis Masqueray.
3. To get to the Statue of Liberty, you can take a ferry from Battery Park in Lower Manhattan. There are several ferry services that offer tours to the statue, and it's important to plan ahead and book your tickets in advance to avoid long lines. Once you're on Liberty Island, you can take an elevator or stairs to the top of the pedestal, where you'll have a stunning view of
4. New York Harbor and the surrounding city skyline. It's important to note that the pedestal is closed to visitors, but you can still get a close-up view of the statue from the observation deck.
5. Be sure to check the weather forecast before your visit, as the statue can be closed due to high winds or other inclement weather conditions. And don't forget to bring your camera to capture the stunning views and memorable moments!

mPLUG-Owl:

As you're looking at a rendering of an iconic white and gold statue of a woman in a field, pay attention to the details of the sculpture, the sky, and the surroundings. The statue is situated in a green field, which provides a picturesque and serene atmosphere. The sky in the background is blue, suggesting a clear and sunny day. To fully appreciate the beauty and significance of the statue, it's crucial to observe its detail, size, and the overall composition of the image. The statue may depict a woman, an important figure in history, or a symbol of liberty and freedom. Additionally, pay attention to the field's surroundings, such as the grass, vegetation, and any other elements that contribute to the overall aesthetic.

LLaVA:

When planning a trip to visit the Statue of Liberty, there are several things you should pay attention to:

1. Tickets: You will need to purchase tickets in advance to visit the Statue of Liberty and Ellis Island. You can do this through the official Statue of Liberty National Monument website or through authorized ticket sellers.
2. Accessibility: The Statue of Liberty is accessible via ferry from Battery Park in Lower Manhattan. There are also guided tours available, which can provide more information about the history and significance of the statue.
3. Closed areas: Some areas of the Statue of Liberty, such as the pedestal and the crown, are closed to visitors due to safety concerns. Make sure to check the official website for the most up-to-date information on accessibility and restrictions.
4. Clothing and footwear: Wear comfortable shoes and dress appropriately for the weather, as you will be walking and standing for extended periods of time.

Figure 10: In this case study, for the first question, EasyGen can give an accurate answer, but the responses of the other two models are a bit biased. For the second question, EasyGen and LLaVA both give reasonable advice.



How can the girl improve her working environment?

EasyGen (ours):

In the image, the young girl is sitting in a room with a laptop in front of her, which suggests that she might be working or studying. To improve her working environment, she can consider several factors:

1. Ergonomics: Ensure that the laptop is placed at eye level or slightly below, and the desk is at a comfortable height for her to maintain a neutral posture. Adjustable furniture like a desk or a chair can help achieve a comfortable working height.
2. Lighting: Ensure that the room is well-lit, with natural light or artificial lighting to minimize eye strain and improve focus.
3. Organization: Keep the room clean and organized, with important items within easy reach. This can help maintain focus and prevent distractions.
4. Comfortable environment: Creating a comfortable environment, such as a dedicated workspace separate from her bedroom, can help establish boundaries and promote productivity.
5. Time management: Establish a routine and allocate specific time slots for studying or working on her laptop. This can help her stay on track and avoid procrastination.

By implementing these suggestions, the girl can enhance her working environment and achieve better focus, productivity, and overall comfort while working on her laptop.

mPLUG-Owl:

To improve her working environment, the girl could consider several options:

1. Organize her space: Ensure the area where she is working is clean, uncluttered, and well-lit. Remove distractions such as toys or other items that might cause her to lose focus. Arrange her books and laptop in a manner that is easy to access and comfortable for her to work in.
2. Create a designated work area: Set up a designated space for her to work, such as a desk, table, or a specific corner in the living room. This helps her maintain her focus and minimize distractions.
3. Provide a comfortable chair and proper lighting: Select a chair with good back support and proper ergonomics to ensure she is comfortable while working. Lighting should be bright and evenly distributed to minimize shadows and glare.
4. Set reasonable work hours: Set a reasonable time limit for her to work, and let her know when it's time to take a break. This can help her maintain focus and minimize fatigue.

By creating a well-organized, comfortable, and productive working environment, the girl can improve her focus and efficiency while minimizing distractions during her time spent studying and doing homework.

MiniGPT-4:

There are several ways the girl can improve her working environment. One way is to make sure her desk and chair are comfortable and adjustable to her height. Another way is to have a clean and organized workspace, so she can easily find what she needs. Additionally, she could have a good lighting system to help her see her work better. Finally, she could have a cushion or pillow to sit on to help reduce fatigue.

InstructBlip:

The girl in the image can improve her working environment by organizing her surroundings and creating a designated study area. To achieve this, she could separate the laptop from other items in the room and create a consistent workspace. This can help her maintain focus and reduce distractions, which can ultimately enhance her productivity and creativity. Additionally, using a comfortable chair and positioning the laptop at a suitable height can help prevent eye strain and promote better posture during long periods of work. Using a calm atmosphere with adequate lighting and placing the laptop away from windows or other sources of glare can also aid in ensuring a more conducive working environment.

Figure 11: From this example, we can find that the response from EasyGen is more comprehensive and coherent. This shows EasyGen can give reasonable suggestions based on the given image.

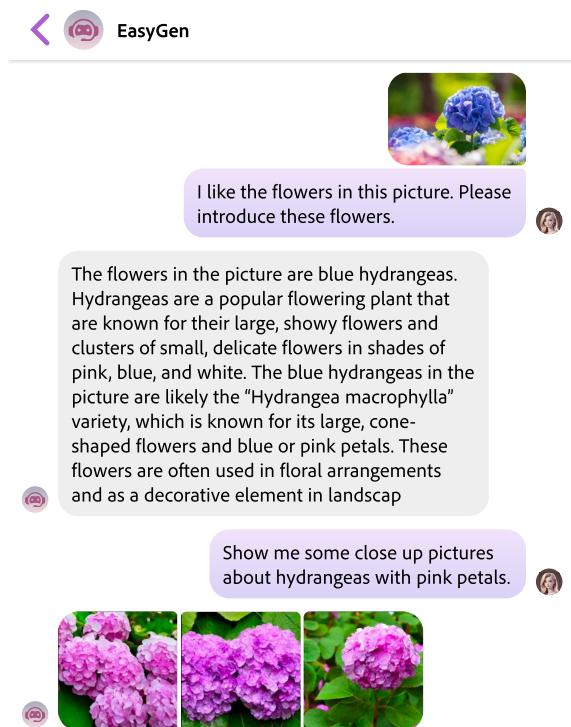


Figure 12: Example of multimodal response generation.

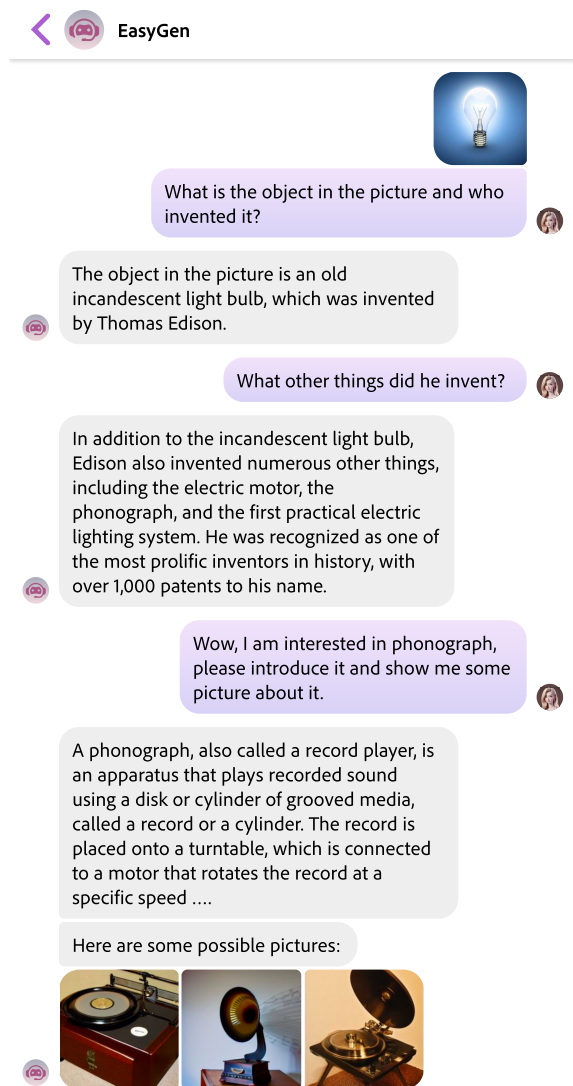


Figure 13: Example of multimodal response generation.