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 the gap between modalities, EasyGen leverages a bidirectional conditional diffusion model named BiDiffuser, which promotes more efficient interactions between modalities. EasyGen handles image-to-text generation by integrating BiDiffuser and an LLM via a simple projection layer. Unlike most existing multimodal models that are limited to generating text responses, EasyGen can also facilitate text-to-image generation by utilizing the LLM to create textual descriptions, which can be interpreted by BiDiffuser to generate appropriate visual responses. Furthermore, EasyGen can be effortlessly integrated into existing advanced multimodal LLMs like LLaVA to improve their performance. Extensive quantitative and qualitative experiments demonstrate the effectiveness of EasyGen, whose training can be easily achieved in a lab setting.

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 large-scale multimodal generative models, 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. Furthermore, most of existing multimodal models lack the ability of generating multimodal responses, as they are primarily designed to understand multimodal content and often fall short in generating content beyond text. A recent work Emu (Sun et al., 2023) takes a unified approach to generate visual or textual tokens, but it also heavily relies on vast quantities of training data.

In this work, we propose EasyGen, a model that facilitates multimodal generation by harnessing the strengths of diffusion models and LLMs. Diffusion models with multiple skills, such as Versatile Diffusion (Xu et al., 2023) and UniDiffuser (Bao et al., 2023b), have exhibited remarkable capabilities in accurately capturing multimodal distributions. UniDiffuser, in particular, shows great potential in accommodating multimodal interactions. It views both image and text as sequential token streams for diffusion calculations, making it well-suited for
Text Generation

As illustrated in Figure 2, we bridge BiDiffuser and To address this limitation, we finetune UniDiffuser accurate visual responses, as illustrated in Figure 2. dialogues, which can aid BiDiffuser in generating

subscriptions and cues derived from text contexts like the LLM can be utilized to generate detailed de-
synchronizing its embedding space with that of an

structured as sequences. However, since UniDif-
fuser aims to fit all conditional distributions (in-

the text representations in LLMs that are typically

structured as sequences. However, since UniDif-
fuser aims to fit all conditional distributions (in-

cluding those conditioned on noisy inputs) into one

model, it is less effective on particular tasks such as conditional generation based on noise-free inputs.

To address this limitation, we finetune UniDiffuser with a specific focus on the targeted image-to-text and text-to-image tasks. The finetuned model, referred to as BiDiffuser, forms a core component of EasyGen for text and image generation.

BiDiffuser is able to convert image data into a textual format, which simplifies the process of

synchronizing its embedding space with that of an LLM for semantic comprehension and reasoning. As illustrated in Figure 2, we bridge BiDiffuser and the LLM using a simple projection layer, which can be trained efficiently with a small amount of data for image-to-text tasks such as image captioning and visual question answering. Alternatively, the LLM can be utilized to generate detailed descriptions and cues derived from text contexts like dialogues, which can aid BiDiffuser in generating accurate visual responses, as illustrated in Figure 2.

Unconditional Generation

Given a data sample taken from a real data distribution $x_0 \sim q(x_0)$, 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:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t | x_{t-1}),$$

(1)

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I),$$

(2)

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

$$p(x_{0:T}) = p(x_T) \prod_{t=0}^{T} p_{\theta}(x_{t-1}|x_t),$$

(3)

$$p_{\theta}(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_t(x_t, t), \sigma^2_t I),$$

(4)

where $p(x_T) = \mathcal{N}(x_T; 0, I)$ is the standard Gaussian distribution and $\mu_t(\cdot)$ is the parameterization of the predicted mean. Diffusion models are trained

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.

Figure 2: Overview of EasyGen.

Figure 1 demonstrates the capability of EasyGen in processing multimodal inputs and generating the appropriate multimodal responses. Notably, EasyGen achieves competitive performance compared to state-of-the-art models with less training data and training time. Without employing parameter-efficient fine-tuning techniques like LoRa (Hu et al., 2021), EasyGen only requires about 120 A100 (80G) GPU hours during the pre-training process (for training BiDiffuser) and 20/72 A100 (80G) GPU hours during the alignment process for FlanT5XL/Vicuna-7B (see Table 10 in Appendix H). Furthermore, EasyGen can be conveniently incorporated into contemporary advanced multimodal LLMs such as LLaVA to improve their performance, as depicted in Figure 6 in Appendix G.

2 Basics of Diffusion Models

Unconditional Generation

Given a data sample taken from a real data distribution $x_0 \sim q(x_0)$, 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:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t | x_{t-1}),$$

(1)

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I),$$

(2)

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

$$p(x_{0:T}) = p(x_T) \prod_{t=0}^{T} p_{\theta}(x_{t-1}|x_t),$$

(3)

$$p_{\theta}(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_t(x_t, t), \sigma^2_t I),$$

(4)

where $p(x_T) = \mathcal{N}(x_T; 0, I)$ 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 $\mathbb{E}[\log p_θ(x_0)]$, and the canonical objective is the variational lower bound of $\log p_θ(x_0)$. Denoising diffusion probabilistic models (Ho et al., 2020) generate samples $x_t \sim q(x_t|x_0)$ by injecting noise $\epsilon \sim N(0, I)$ to the data $x_0$, and train a network $\epsilon_t(\cdot)$ to predict the added noise $\epsilon$ using a standard mean squared error loss:

$$ L := \mathbb{E}_{x_0, \epsilon_t} [\| \epsilon - \epsilon_t(x_t, t) \|^2]. \quad (5) $$

Note that $\mu_t(x_t, t)$ can be derived from $\epsilon_t(x_t, t)$.

**Conditional Generation** For conditional generation, a paired data $(x_0, y_0)$ is given, and the aim is to model the conditional data distribution $q(x_0|y_0)$, where $y_0$ 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 the sample quality. For classifier-free guidance, no classifier needs to be trained. The denoising network $\epsilon_t(x_t|y_0)$ simply conditions on the information encoded in $y_0$. At inference time, with a guidance scale $s$, the modified score estimate is further in the direction of $\epsilon(x_t|y_0)$ and away from the unconditional model $\epsilon_t(x_t|\emptyset)$ ($\emptyset$ is a null token) as follows:

$$ \tilde{\epsilon}_t(x_t|y_0) = \epsilon_t(x_t|\emptyset) + s \cdot (\epsilon_t(x_t|y_0) - \epsilon_t(x_t|\emptyset)). \quad (6) $$

### 3 EasyGen: Easy Multimodal Generation with a Bidirectional Conditional Diffusion Model and LLMs

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.

#### 3.1 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 (Bao et al., 2023b), 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 $(x_0, y_0)$ and generates noisy data $x_{t^x}$ and $y_{t^y}$, where $0 \leq t^x, t^y \leq T$ represent two individual timesteps (perturbation levels). It then trains a joint noise prediction network $\epsilon_t(x_{t^x}, y_{t^y}, t^x, t^y)$ to predict the noise $\epsilon^x$ and $\epsilon^y$ by minimizing the mean squared error loss:

$$ \mathbb{E}_{\epsilon^x, \epsilon^y, x_0, y_0} [\| \epsilon^x - \epsilon_t(x_{t^x}, y_{t^y}, t^x, t^y) \|^2], \quad (7) $$

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

By predicting $\epsilon_t(x_{t^x}, y_{t^y}, t^x, t^y)$ for any $t^x$ and $t^y$, UniDiffuser learns all distributions related to $(x_0, y_0)$ simultaneously. This includes all conditional distributions: $q(x_0|y_0)$ for text-to-image generation, $q(y_0|x_0)$ for image-to-text generation, and those conditioned on noisy input, i.e., $q(x_0|y_{t^y})$ and $q(y_0|x_{t^x})$, for $0 < t^x, t^y \leq T$. Learning a conditional distribution $q(x_0|y_{t^y})$ or $q(y_0|x_{t^x})$ 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 $q(x_0|y_0)$ and $q(y_0|x_0)$.

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:

$$ L = \mathbb{E}_{\epsilon^x, \epsilon^y, x_0, y_0} [\| \epsilon^x - \hat{\epsilon}_t(x_{t^x}, y_0, t^x, 0) \|^2 + \alpha \| \epsilon^y - \hat{\epsilon}_t(x_0, y_{t^y}, 0, t^y) \|^2]. \quad (8) $$

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 $y_0$ based on the input image $x_0$ and vice versa, where the input conditions for the model are noise-free. We employ classifier-free guidance. During training, we estimate the noise injected to the image (i.e., $\hat{\epsilon}_t^x(x_{t^x}, y_0, t^x, 0)$) conditioned on the noise-free text $y_0$ and the noise to the text (i.e., $\hat{\epsilon}_t^y(x_0, y_{t^y}, 0, t^y)$) given the noise-free image $x_0$. During inference, with a guidance scale $s \geq 0$, we use the modified prediction $\hat{\epsilon}_t$ to guide towards the
3.2.1 Aligning BiDiffuser with LLMs

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 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 autoregressive loss:

$$
\mathcal{L}_{\text{ITG}} = -\frac{1}{L} \sum_{l=1}^{L} \log p(y^g_{l} | w^g_{<l}, I, T), \quad (10)
$$

where $w^g = (w^g_1, \ldots, w^g_L)$ represents the ground-truth caption of image $I$ with length $L$, $T$ is the text instruction, and $\theta$ 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 $d_{\text{diff}}$, should align with the image caption that is encoded by the LLM’s encoder, denoted as $d_{\text{lin}}$. Therefore, to accurately learn the alignment between the image and text representations, in addition to the ITG loss in Eq. 10, we also employ an image-text distance minimization (ITDM) loss:

$$
\mathcal{L}_{\text{ITDM}} = \frac{1}{N} \sum_{i=1}^{N} \|d_{\text{diff}} - d_{\text{lin}}\|_{2}^{2}, \quad (11)
$$

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 aligning BiDiffuser with LLMs, EasyGen gains the capability of zero-shot image-to-text generation, which includes tasks such as image captioning and VQA.

### 3.2.2 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 1. 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 (blue texts in Table 9), which enable the LLM to produce concise responses when necessary. For different tasks, the distinct instruction templates are as outlined in Appendix G.

### 3.3 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.

To tackle multimodal response generation tasks such as PhotoChat (Zang et al., 2021), we first fine-tune the LLM 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 + <Img><caption></Img> + Dialog history. **Assistant:** <response>.

**USER:** Dialog history. **Assistant:** <response>.

![Diagram](image-url)

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

<table>
<thead>
<tr>
<th>Different Instruction Templates:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Caption</strong></td>
</tr>
<tr>
<td><strong>Generation</strong></td>
</tr>
<tr>
<td><strong>VQA</strong></td>
</tr>
</tbody>
</table>

| **Multimodal Dialogue** | **Response** | Table 1: Examples of different instructions with different output formats. We use bold fonts to indicate different instructions. |
We initialize the encoder-decoder LLM using pre-trained weights from FlanT5XL 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 C. 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, our method is 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 layer on the PhotoChat dataset. We incorporate photogenerating 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 J, we regenerate image annotations in a text format similar to that used in MS-COCO.

### 4 Experiments

#### 4.1 Experimental Setup

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 D and F.

#### 4.2 Evaluation

Table 2 lists the evaluation results for each baseline and our models on MS-COCO and VQA (zero-shot) datasets. EasyGen outperforms most of the baseline models on both the COCO test set and NoCaps test set (zero-shot transfer). Although EasyGen is only pre-trained on a small dataset MS-COCO, its performance on the image captioning generation task is comparable to models (e.g., BLIP-2) pre-trained on a large dataset. This indicates that EasyGen can effectively combine the strength of diffusion module and LLM to generate smooth and informative captions. GPT scores do not vary significantly because the captions produced by the models in the image-captioning task tend to be quite alike. For the OK-VQA and GQA dataset, the performance of EasyGen is improved compared with other models of a similar scale. For example, BLIP-2 adopts the task-special decoding method and achieves 39.4% accuracy on OK-VQA validation set, while ours can get 45.2% even with a simple decoding method, i.e., greedy search.

Table 3 lists the evaluation results on the PhotoChat dataset. The results of Divter are cited from (Sun et al., 2021). We fine-tune Maria on PhotoChat dataset only for the response generation task. Since our EasyGen model can generate response and image description simultaneously, the response and description generation task has a similar PPL. Compared with other models, our method has clear advantages in the performance of PPL, indicating that our model demonstrates strong performance on text generation tasks. Besides, we find that the image descriptions in the PhotoChat dataset are too concise to convey the information of images.

#### 4.3 Overall Results

Table 2 lists the evaluation results for each baseline and our models on MS-COCO and VQA (zero-shot) datasets. EasyGen outperforms most of the baseline models on both the COCO test set and NoCaps validation set (zero-shot transfer). Although EasyGen is only pre-trained on a small dataset MS-COCO, its performance on the image captioning generation task is comparable to models (e.g., BLIP-2) pre-trained on a large dataset. This indicates that EasyGen can effectively combine the strength of diffusion module and LLM to generate smooth and informative captions. GPT scores do not vary significantly because the captions produced by the models in the image-captioning task tend to be quite alike. For the OK-VQA and GQA dataset, the performance of EasyGen is improved compared with other models of a similar scale. For example, BLIP-2 adopts the task-special decoding method and achieves 39.4% accuracy on OK-VQA validation set, while ours can get 45.2% even with a simple decoding method, i.e., greedy search.

Table 3 lists the evaluation results on the PhotoChat dataset. The results of Divter are cited from (Sun et al., 2021). We fine-tune Maria on PhotoChat dataset only for the response generation task. Since our EasyGen model can generate response and image description simultaneously, the response and description generation task has a similar PPL. Compared with other models, our method has clear advantages in the performance of PPL, indicating that our model demonstrates strong performance on text generation tasks. Besides, we find that the image descriptions in the PhotoChat dataset are too concise to convey the information of images.
Table 2: Automatic evaluation and GPT evaluation of our model and the baselines on various vision-language tasks. PT, IT indicate the number of samples in pretraining and instruction tuning stage respectively. The results of EasyGen on NoCaps, OK-VQA and GQA are obtained in a zero-shot setting. † indicates that the model was trained on other VQA datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset Size</th>
<th>NoCaps (val)</th>
<th>COCO (Karpathy)</th>
<th>OK-VQA Accuracy</th>
<th>GQA Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLIP (Li et al., 2022)</td>
<td>129M</td>
<td>-</td>
<td>113.2</td>
<td>40.4</td>
<td>-</td>
</tr>
<tr>
<td>Flamingo (Alayrac et al., 2022)</td>
<td>1.8B</td>
<td>-</td>
<td>14.8</td>
<td>138.1</td>
<td>-</td>
</tr>
<tr>
<td>BLIP-2 OPT-6.7B (Li et al., 2023b)</td>
<td>129M</td>
<td>-</td>
<td>43.5</td>
<td>145.2</td>
<td>-</td>
</tr>
<tr>
<td>BLIP-2 FlanT5XL (Li et al., 2023b)</td>
<td>129M</td>
<td>-</td>
<td>15.8</td>
<td>42.4</td>
<td>-</td>
</tr>
<tr>
<td>InstructBlip 7B (Dai et al., 2023)</td>
<td>1.2M</td>
<td>-</td>
<td>43.5</td>
<td>145.2</td>
<td>-</td>
</tr>
<tr>
<td>MiniGPT-4 (Zhu et al., 2023)</td>
<td>-</td>
<td>5M</td>
<td>121.6</td>
<td>37.5</td>
<td>30.8</td>
</tr>
<tr>
<td>LLaVA (Liu et al., 2023b)</td>
<td>558K</td>
<td>158K</td>
<td>123.1</td>
<td>-</td>
<td>61.0*</td>
</tr>
<tr>
<td>EasyGen FlanT5XL</td>
<td>169K</td>
<td>90K</td>
<td>121.2</td>
<td>42.4</td>
<td>41.1</td>
</tr>
<tr>
<td>EasyGen Vicuna-7B</td>
<td>169K</td>
<td>90K</td>
<td>121.8</td>
<td>44.6</td>
<td>44.4</td>
</tr>
</tbody>
</table>

Table 3: Automatic evaluation of our model and the baselines on the PhotoChat dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Response Generation</th>
<th>Description Generation</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divter (Sun et al., 2021)</td>
<td>6.5</td>
<td>15.1/11.4</td>
<td>29.16</td>
</tr>
<tr>
<td>Maria (Liang et al., 2021)</td>
<td>13.8</td>
<td>15.8/13.2</td>
<td>-</td>
</tr>
<tr>
<td>EasyGen FlanT5XL</td>
<td>22.3</td>
<td>13.5/10.2</td>
<td>4.16</td>
</tr>
<tr>
<td>+ w/ generated desc.</td>
<td>17.8</td>
<td>17.4/13.2</td>
<td>6.23</td>
</tr>
<tr>
<td>+ w/o generated desc.</td>
<td>18.7</td>
<td>4.16</td>
<td>75.46</td>
</tr>
</tbody>
</table>

Therefore, we used the pre-trained model from the first stage to regenerate the image description (referred to as “w/ generated desc.” in Table 3) which led to a large gap towards ground-truth descriptions, resulting in lower BLEU-1/2 and ROUGE-L. However, the performance of our model on BLEU-1/2 and ROUGE is higher than other models on response generation tasks, indicating that introducing richer image descriptions is beneficial for generating more relevant and informative responses.

4.4 Ablation Study

In Table 4, we examine the impact of freezing/tuning BiDiffuser and the LLM. We conducted ablation studies on image captioning and VQA tasks. It can be observed that the frozen Mid-Align method outperforms the Pre-Align method in image captioning. This shows that the 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 TS and Vicuna. Compared to models at the same scale, these instruction-tuned models achieve superior results.

4.5 Fine-tuning for VQA Tasks

In order to enhance the performance of EasyGen on VQA and OCR tasks, we take into account the potential information dilution or omission that may occur when using BiDiffuser to convert images to text space. So we opt to integrate BiDiffuser with CLIP ViT-L/14 as image encoder (as shown in Figure 6 in Appendix). In this process, we freeze CLIP and BiDiffuser while fine-tuning the parameters of the LLM and projection layers. Table 5 presents the results on traditional short QA and the modern benchmark MMBench (Liu et al., 2023c). For fine-tuning EasyGen, we utilize train and val splits from VQAv2, Text Captions, AOK-VQA, and TextVQA datasets. Though we do not pre-align CLIP with LLM and use smaller instruction data for tuning, EasyGen’s performance is better than LLaVA on VQAv2 dataset, which shows BiDiffuser can effectively help LLM understand images. Furthermore, BiDiffuser can also be plug-and-played into other MLLMs and improve their performance. We speculate that BiDiffuser provides guidance information to MLLMs, enabling them to better understand the details of CLIP encoded images.

5 Related Work

Multimodal Language Models. Recent research has witnessed a surge of interest in multimodal LLMs, including collaborative models...
such as Visual ChatGPT (Wu et al., 2023a), MM-REACT (Yang et al., 2023), and Hugging-GPT (Shen et al., 2023), and end-to-end methods including Flamingo (Alayrac et al., 2022), Img2LLM (Guo et al., 2022), BLIP series (Li et al., 2023b; Dai et al., 2023; Li et al., 2022), BEiT series (Bao et al., 2021; Wang et al., 2022b), LLava (Liu et al., 2023b), mPLUG-owl (Ye et al., 2023), MiniGPT-4 (Zhu et al., 2023), Llama-adapter (Zhang et al., 2023a), Otter (Li et al., 2023a), OFA (Wang et al., 2022a), and PaLI (Chen et al., 2022), EasyGen is built upon a bidirectional conditional diffusion model, 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 works have shown significant progress on multimodal response generation (Koh et al., 2023b; Aghajanyan et al., 2022; Zhang et al., 2023b; Wu et al., 2023b; Pan et al., 2023; Koh et al., 2023a). Divter (Sun et al., 2021) incorporates text-to-image generation into text-only dialogue response generation to produce a multimodal response. Leveraging the power of diffusion models, CoDi (Tang et al., 2023) can generate any combination of output modalities. Emu (Sun et al., 2023) takes a unified approach to predict the subsequent visual or textual token in an auto-regressive manner. In EasyGen, we efficiently combine the diffusion model and LLMs to generate multimodal outputs.

**6 Conclusion**

We have introduced EasyGen, a model that facilitates multimodal understanding and generation. In contrast to existing models that rely on encoders like CLIP or ImageBind and require significant amounts of training data to integrate different modalities, EasyGen offers a more efficient solution by employing a bidirectional diffusion model named BiDiffuser. 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.
References


Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructclip: Towards general-purpose vision-language models with instruction tuning.


A 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-captions 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.

B 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.

C Datasets

We test the effectiveness of EasyGen by experimenting on different tasks including image captioning, visual question answering (VQA), and multimodal dialogue tasks. Table 6 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 dialogue context, and the subsequent turns as the prediction targets. By converting the dialogues of this dataset into the form mentioned in 3.3, 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 8) 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.

D 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 Clinghilla 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 gen-
Table 6: Description of datasets used in our alignment and VQA fine-tuning stages. Noting that in alignment process, we used 5K images from VQA2 dataset.

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

Table 7: Summary of the evaluation datasets and metrics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Split</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-COCO (Lin et al., 2014)</td>
<td>Image captioning</td>
<td>Test</td>
<td>CIDEr, BLEU, SPICE, Accuracy</td>
</tr>
<tr>
<td>NoCaps (Agrawal et al., 2019)</td>
<td>Image captioning</td>
<td>Val</td>
<td>CIDEr, SPICE, Accuracy</td>
</tr>
<tr>
<td>OK-VQA (Marino et al., 2019)</td>
<td>Image captioning</td>
<td>Val</td>
<td>CIDEr, SPICE, Accuracy</td>
</tr>
<tr>
<td>GQA (Hudson and Manning, 2019)</td>
<td>VQA</td>
<td>Test</td>
<td>Accuracy, Accuracy</td>
</tr>
</tbody>
</table>

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 pretrained 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 conversation 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.

**E 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.

**F Implementation Details**

LLM During the alignment process, we utilize the AdamW optimizer with $\beta_0 = 0.9$, $\beta_1 = 0.99$, $\epsilon = 1e-8$, and a learning rate of $1e-5$. The batch size is set to 128 images, and the model is trained for 5 epochs. The weight decay is set to 0.01, and the AdamW optimizer is used.
Table 8: 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.

<table>
<thead>
<tr>
<th>Task</th>
<th>Instruction Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Captioning</td>
<td>USER: &lt;image&gt;+random[query] Assistant:</td>
</tr>
<tr>
<td>LLaVA 80K</td>
<td>USER: Please answer question from this image: &lt;image&gt; Question: &lt;question&gt; Assistant: USER: Answer question &lt;question&gt; through the image &lt;image&gt; Assistant:</td>
</tr>
<tr>
<td>Multimodal Dialogue</td>
<td>USER: Dialog history+&lt;photo&gt;+Dialogue history Assistant:</td>
</tr>
<tr>
<td>VQA</td>
<td>USER: Image: &lt;image&gt; Question: &lt;question&gt; Short answer: Assistant:</td>
</tr>
</tbody>
</table>
### Table 10: EasyGen’s trainable parameters, training data size, and training cost during alignment process.

<table>
<thead>
<tr>
<th>Model</th>
<th>Trainable Param.</th>
<th>Training Images</th>
<th>Training Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiDiffuser</td>
<td>952M</td>
<td>169K</td>
<td>120 (A100 80GB) GPU hours</td>
</tr>
<tr>
<td>Alignment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Projection Layers + TSXL</td>
<td>4M</td>
<td>163K</td>
<td>20 (RTX3090 24GB) GPU hours</td>
</tr>
<tr>
<td>Projection Layers + TSXL</td>
<td>3B</td>
<td>173K</td>
<td>20 (A100 80GB) GPU hours</td>
</tr>
<tr>
<td>Projection Layers + Vicuna 7B</td>
<td>7B</td>
<td>173K</td>
<td>72 (A100 80GB) GPU hours</td>
</tr>
<tr>
<td>Projection Layers + Vicuna 7B(LoRa)</td>
<td>610M</td>
<td>251K</td>
<td>20 (A100 80GB) GPU hours</td>
</tr>
</tbody>
</table>

### Table 11: Ablation studies on the instruction-tuning process and loss functions.

<table>
<thead>
<tr>
<th>Model</th>
<th>NoCaps (val)</th>
<th>COCO (Karpathy)</th>
<th>OK-VQA</th>
<th>GQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>EasyGen Mid-Align FlanTSXL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ w/o ITDM</td>
<td>118.6</td>
<td>15.3</td>
<td>24.8</td>
<td>42.2</td>
</tr>
<tr>
<td>+ w/o ITG</td>
<td>93.2</td>
<td>12.9</td>
<td>23.0</td>
<td>35.1</td>
</tr>
</tbody>
</table>

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.

Figure 8: Examples of image captioning results by EasyGen.
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.

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.

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.

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.

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:

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.

Here are some pictures that resemble the painting, featuring young girls with their hair up in braids and wearing gold jewelry: 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.
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.

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 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.

4. 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!

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 latest 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.
How can the girl improve her working environment?

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.

---

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.
EasyGen

I like the flowers in this picture. Please introduce these flowers.
Show me some close up pictures about hydrangeas with pink petals.

The flowers in the picture are blue hydrangeas. Hydrangeas are a popular flowering plant that are known for their large, showy flowers and clusters of small, delicate flowers in shades of pink, blue, and white. The blue hydrangeas in the picture are likely the “Hydrangea macrophylla” variety, which is known for its large, cone-shaped flowers and blue or pink petals. These flowers are often used in floral arrangements and as a decorative element in landscapes.

EasyGen

What is the object in the picture and who invented it?
What other things did he invent?
Wow, I am interested in phonograph, please introduce it and show me some picture about it.

The object in the picture is an old incandescent light bulb, which was invented by Thomas Edison.
In addition to the incandescent light bulb, Edison also invented numerous other things, including the electric motor, the phonograph, and the first practical electric lighting system. He was recognized as one of the most prolific inventors in history, with over 1,000 patents to his name.

A phonograph, also called a record player, is an apparatus that plays recorded sound using a disk or cylinder of grooved media, called a record or a cylinder. The record is placed onto a turntable, which is connected to a motor that rotates the record at a specific speed ….

Here are some possible pictures:

Figure 12: Example of multimodal response generation.
Figure 13: Example of multimodal response generation.