# De-Diffusion Makes Text a Strong Cross-Modal Interface

Chen Wei[1,2] Chenxi Liu[1] Siyuan Qiao[1] Zhishuai Zhang[1] Alan Yuille[2] Jiahui Yu[1]
[1]Google DeepMind [2]Johns Hopkins University

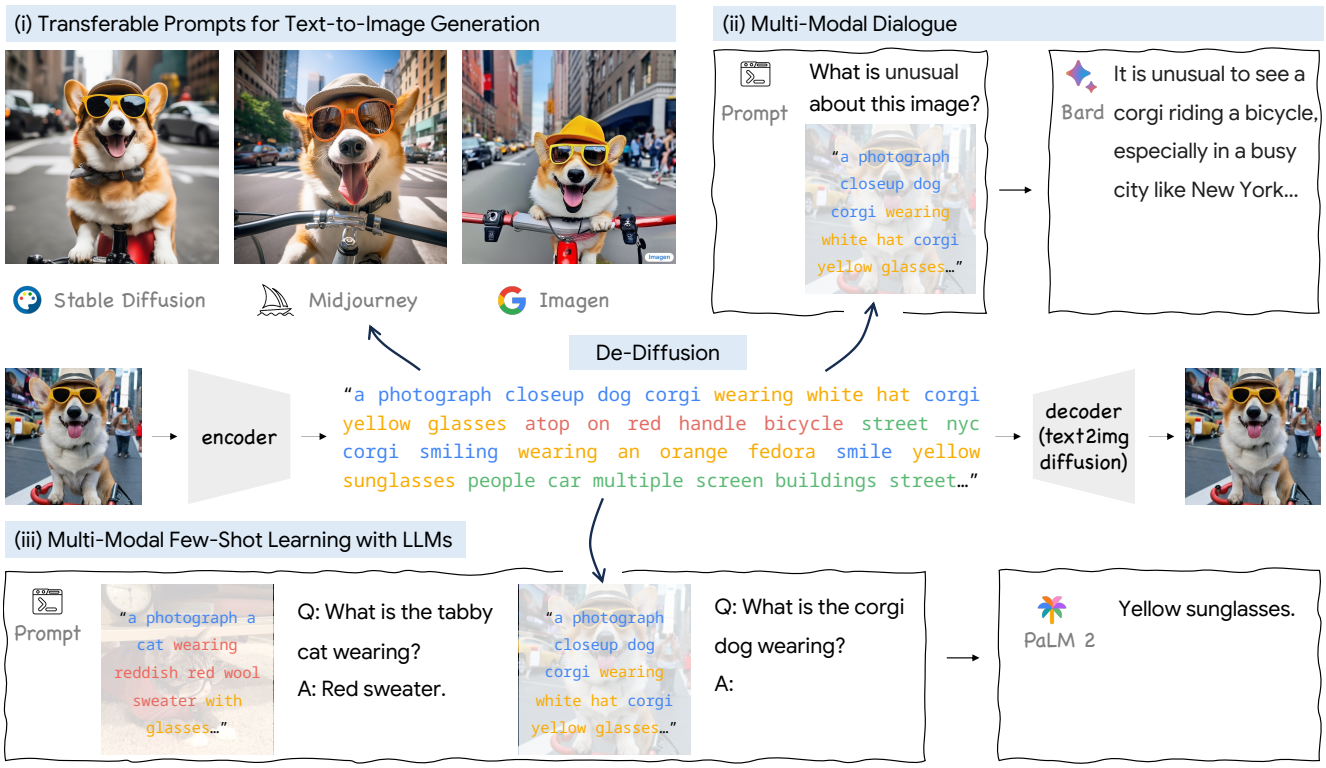


Figure 1. De-Diffusion is an autoencoder whose decoder is a pre-trained text-to-image diffusion model. It encodes an input image into a piece of information-rich text, which mixes comprehensive semantic concepts present in the image to be a "scrambled caption". We group semantics by color for illustration. De-Diffusion text can act as a flexible interface between different modalities, for example, enabling diverse vision-language applications including: (*i*) providing transferable prompts for different text-to-image tools, (*ii*) enabling text-only chatbots, *e.g.*, Bard [1], to engage in multi-modal dialogue, and (*iii*) injecting image context into off-the-shelf large language models (LLMs), *e.g.*, PaLM 2 [5], to perform open-ended visual question answering by prompting the LLM with few-shot examples.

## Abstract

*We demonstrate text as a strong cross-modal interface. Rather than relying on deep embeddings to connect image and language as the interface representation, our approach represents an image as text, from which we enjoy the interpretability and flexibility inherent to natural language. We employ an autoencoder that uses a pre-trained text-to-image diffusion model for decoding. The encoder is trained to transform an input image into text, which is then fed into the fixed text-to-image diffusion decoder to reconstruct the original input – a process we term* De-Diffusion. *Experiments validate both the precision and comprehensiveness of De-Diffusion text representing images, such that it can be readily ingested by off-the-shelf text-to-image tools and LLMs for diverse multi-modal tasks. For example, a single De-Diffusion model can generalize to provide transferable prompts for different text-to-image tools, and also achieves a new state of the art on open-ended vision-language tasks by simply prompting large language models with few-shot examples. Project page: dediffusion.github.io.*

## 1. Introduction

We have witnessed LLM-powered products such as ChatGPT taking over the world by storm. Nowadays many people are convinced of the benefits that LLMs can bring in understanding natural language conversations and assisting humans in creative tasks. However, what is the path forward? One clear direction and trend is towards *multi-modality*, allowing the model to understand additional modalities such as image, video, and audio. GPT-4 [57] is a multi-modal model with impressive image understanding capabilities, and has recently rolled out to the public together with audio-processing capabilities. Gemini is also "multi-modal from day one" [2]. Multi-modal models like these have a fundamental design choice to make, *i.e.*, how different modalities should communicate and connect? In the context of this work, we rephrase the question as: what is the *cross-modal interface*?

We argue that a good cross-modal interface should at least possess the following two properties: (1) *content preserving*, *i.e.*, signals from the original modality can be reconstructed from the interface representation to a high degree; (2) *semantically meaningful*, *i.e.*, the interface representation contains useful abstractions of the raw signals, so that understanding and reasoning can be performed more easily. Balancing these two properties is challenging, and in fact they can often be in contention with each other. For example, the raw signals from the original modality satisfy content preserving perfectly, but are lacking on the semantically meaningful front.

Ever since the deep learning era [18, 31, 32, 44], *deep embeddings* have been the go-to choice as cross-modal interface. They can be good at preserving image pixels if trained as an autoencoder [32], and can also be semantically meaningful, with the most recent exemplar being CLIP [59]. In this paper, we do not argue that deep embeddings are a bad cross-modal interface *per se*, but instead convey the idea that according to our experiments, *text* can be a strong alternative cross-modal interface.

If we consider the relationship between the speech and text for a quick second, text has always been so natural of a cross-modal interface that we do not typically think of it as such. Converting the speech audio to text well *preserves the content* such that we can reconstruct the speech audio with the mature text-to-speech technique. We are also confident that the transcribed text contains all the semantics information, in other words, *semantically meaningful*. By analogy, we can also "transcribe" an image into text, which has the more familiar name of image captioning. But when we compare typical image captions against the two properties of cross-modal interface, they do not preserve content well but only capture the most salient semantic concepts. In other words, image captions are more about precision than comprehensiveness [13, 83], and it is hard to answer any and all visual questions from the short captions.

While image captions do not make an ideal interface representation, we argue that precise *and* comprehensive text, if attainable, remains a promising option, both intuitively and practically. Intuitively, humans rely on language to articulate our physical surroundings, engage in reasoning, and deliver solutions. In other words, we constantly "transcribe" information about the external world into language and use it as an interface for higher-level cognition [16, 23]. Practically, text is the native input domain for LLMs. Using text as the interface can avoid the need for adaptive training often required with deep embeddings [4, 46]. Given that training and adapting top-performing LLMs can be prohibitively expensive [4, 5, 57], text provides a modular design that opens up more possibilities. The question is, how can we attain precise and comprehensive text of images?

We resort to the classic autoencoding for a solution [32]. Unlike common autoencoders, we utilize a pre-trained text-to-image diffusion model as the decoder, and naturally, with text as the latent space. The encoder is trained to transform an input image into text, which is then fed into the text-to-image diffusion model for decoding. To minimize the reconstruct error, the latent text, though often mixing semantic concepts together to be a "scrambled caption" of the input image, has to be both precise and comprehensive. No extra supervision is used other than images themselves.

Recent generative text-to-image models excel at converting arbitrary rich text of, *e.g.*, tens of words, to highly detailed images that closely follow the prompts [56, 61, 64, 67, 86]. This essentially suggests the remarkable capability of these generative models to process complex text into visually coherent outputs. By employing one of these generative text-to-image models as the decoder, the optimized encoder explores the wide latent space of text and unpacks the enormous visual-language knowledge encapsulated within the generative model, embodying a foundational paradigm known as Analysis by Synthesis [7, 10, 89].

We show De-Diffusion text extensively captures semantic concepts in images, and, when used as text prompts, enables diverse vision-language applications (Fig. 1). De-Diffusion text can generalize to be a transferable prompt for different text-to-image tools. Evaluated quantitatively by reconstruction FID [30], De-Diffusion text significantly outperforms COCO captions [48] as prompts to a third-party text-to-image model [64]. De-Diffusion text also enables off-the-shelf LLMs to conduct open-ended vision-language tasks by simply prompting LLMs with few-shot task-specific examples. We highlight De-Diffusion outperforms Flamingo [4] on open-ended few-shot VQA [6] with 100$\times$ fewer learnable weights and without using interleaved image-text supervision. The results demonstrate De-Diffusion text effectively interconnects both human interpretations and various off-the-shelf models across domains.

## 2. Related Work

**Autoencoding** is a classical approach for learning representations [32, 66]. It uses an encoder to map the input into a compressed, meaningful representation, and a decoder to reconstruct the input from this representation to be as close as possible to the original. This simple autoencoding concept underpins many unsupervised representation learning algorithms across domains [19, 29, 33, 43, 78]. By forcing the model to compress then reconstruct the input, autoencoders discover useful structural representations of the data. For example, Neural De-Rendering [82] is a generalized autoencoder that utilizes a deterministic rendering function as the decoder and maps images into structured and disentangled scene descriptions. Inspired by its name "de-rendering", we name our approach "De-Diffusion".

A specific type of autoencoder, VQ-VAE [63, 75] or discrete VAE [62], is designed to learn discrete, structured representations in the latent space. This can be especially useful for modeling data with categorical or symbolic attributes. These methods are now widely adopted in multimodal models to tokenize images [21, 62, 64, 86]. However, VQ-VAE's latent space is hidden and often entangled, requiring adaptive fine-tuning for downstream tasks. De-Diffusion also utilizes a discrete latent space. In contrast, we directly encode images into a sequence of text, which is directly interpretable. SPAE [87] and LQAE [51] are two recent approaches that encode images into the vocabulary space of a fixed LLM. They jointly learn the encoder and decoder from scratch. Consequently, although the latent space is discrete text, it tends to act as a "cipher code" that only the co-trained decoder can interpret. This limits generalization to human understanding and off-the-shelf LLMs and text-to-image models. In contrast, De-Diffusion utilizes a pre-trained text-to-image diffusion model as the decoder, obtaining interpretable text as the latent representation.

**How many words is an image worth?** The adage "a picture is worth a thousand words" means that still images can convey complex and sometimes multiple ideas more effectively than a mere verbal description. The question, how many words is an image worth, is constantly explored by the computer vision community [22, 24, 25, 49]. For example, "An image is worth 16×16 words", or ViT [20], proposes to take the image patches as tokens (words) and process these tokens by Transformers [76], which has become one of the standard vision backbones now. In this sense, our work can also been seen as "An image is worth 75 words", for we encode input images into a sequence of 75 tokens.

Several prior works also explore to use text to represent images [9, 84] and combine with LLMs. However, these works rely on multiple captioning and classification models, whose outputs are concatenated to be the text representation. Their performance is heavily dependent on the captioning and classification models, and we demonstrate in Sec. 4 that even human-annotation COCO captions can lack the extensive details covered in De-Diffusion text.

**Vision-language models.** The breakthrough in NLP [11, 19, 36, 37, 57, 60, 80], especially their abilities to do few-shot learning, has inspired a large body of vision-language work. A family of vision-language models is based on contrastive learning [28], where images and text are projected in to a same embedding space [40, 42, 47, 58, 59, 85, 90]. De-Diffusion differs from contrastive models as we encode image as text, instead of deep embeddings. Another family of vision-language models fuses vision and language models by jointly training them with large-scale image-text data [4, 14, 45, 50, 55, 62, 85, 88]. In contrast, De-Diffusion takes a modular design with text as the representation, bypassing the heavy cost image-text data collection and jointly training large-scale vision and language models.

## 3. Method

### 3.1. De-Diffusion for Text Representation

**Autoencoder.** Autoencoding is one of the classical methods for representation learning [32, 66]. An autoencoder first encodes an input $x$ into a latent representation $z$, then decodes $z$ back to $\tilde{x}$ for reconstruction. Both the encoder and the decoder are optimized so that the reconstructed input $\tilde{x}$ is as similar as possible to the original input $x$. By doing so, the compressed representation $z$ preserves the information in the input. Since no more supervision is required except the input itself, autoencoding is an unsupervised approach without the heavy burden of human annotation.

**Text as the latent representation.** While autoencoders can learn compressed representations $z$ that preserve useful information, it is difficult to use the latent $z$ for downstream tasks without any additional training, let alone direct human interpretation. In this work, we propose to encode the input image into *text*. Practically, the encoder compresses each image into a sequence of BPE-encoded text tokens [69], where each token can take on a discrete value from the vocabulary. To faithfully reconstruct the image from the latent text, the text must precisely and comprehensively capture the semantic concepts present in the image, making a interface representation, in contrast to image captions that only focus on the most visually salient information.

**Text-to-image diffusion as the decoder.** One potential concern is that the encoder might still encrypt the images into a cipher code that only the decoder can decipher, making human interpretation challenging. This is particularly likely when the encoder and the decoder are jointly trained. To mitigate this concern [51], we introduce a pre-trained text-to-image diffusion model as the decoder, and dub our method as "De-Diffusion".

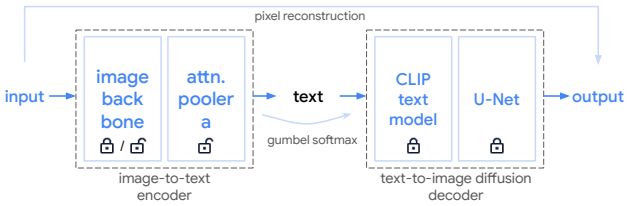


Figure 2. **Architecture of De-Diffusion.** The overall structure is an autoencoder, with (*i*) a pre-trained text-to-image diffusion model as the decoder, (*ii*) text as the latent representation, and (*iii*) a image-to-text encoder. Lock and unlock denote frozen and learnable weights, respectively. We use Gumbel-softmax [41, 53] for discrete text tokens.

Text-to-image diffusion models, as the name suggested, learn the relationship between text and images from a large dataset of image-text pairs and excel at converting texts into highly detailed images. They already establish the projection from descriptive text to image, and we unpack this encapsulated knowledge by employing a frozen text-to-image diffusion model as the decoder. As illustrated in Fig. 2, the text-to-image diffusion model consists of a CLIP text encoder [59] and a U-Net [65], and the codebook is then naturally the vocabulary of the CLIP text encoder.

When training De-Diffusion, we freeze the parameters of the text-to-image diffusion decoder. In each mini-batch, we expose the decoder with one randomly sampled noise level for each sample. This resembles the training procedure for diffusion models [34], except the parameters are fixed and the text conditions are outputs of the image-to-text encoder instead of the training data.

**Image-to-text encoder.** The encoder maps the input image into text. It starts with an image backbone that extracts image features, followed by an attentional pooler [38, 85] that turns the features into output text tokens. The image backbone can be a pre-trained and frozen model that excels at image feature extraction. It can also be randomly initialized, supervised by the reconstruction objective during De-Diffusion training. We ablate the two choices in Tab. 4d.

The attentional pooler projects $n$ learnable queries to $n$ text tokens by a few Transformer blocks [76]. Each Transformer block consists of a self-attention layer over all the queries, a cross-attention layer to gather features from the image backbone, and an MLP layer. After the Transformer blocks, a linear layer projects the queries to discrete text tokens from the vocabulary of CLIP text encoder, in order to connect to the diffusion decoder. The $n$ queries are positional sensitive, meaning that each query corresponds to a specific position in the CLIP text encoder. The $n$ output text tokens, together with the special tokens `[SOS]` and `[EOS]`, are then fed into the diffusion decoder. We ablate the effect of $n$, the number of text tokens, in Tab. 4a.

**Optimization.** Same as other autoencoders, the training objective of De-Diffusion is to minimize the reconstruction error between the input image and the reconstruction from the pre-trained diffusion model. Practically, both the loss function and the noise variance schedule strictly follow those of the decoder, *i.e.*, the pre-trained diffusion model [34]. The training data of De-Diffusion only includes images, without human annotations or text descriptions.

Our model can be viewed as a special discrete autoencoder with discrete text tokens as the latent. Similar to other discrete autoencoders [62, 63, 75], we use Gumbel-softmax [41, 53] as the continuous relaxation to back-propagate the gradients from the decoder through the discrete latent. The relaxation becomes tight as the temperature $\tau \to 0$. We find that an annealing schedule of temperature $\tau$ is important for stable training.

To increase the information density and readability, we exclude all the punctuation in the vocabulary, which accounts for around 6% of the original vocabulary of CLIP text encoder. As a result, only word tokens and number tokens are allowed. We ablation this design choice in Tab. 4b.

### 3.2. Implementation Details

**Text-to-image diffusion model.** The text-to-image diffusion model used for De-Diffusion training is based on Imagen [67]. The U-Net has 600M parameters with an embedding dimension of 256 and input resolution of 64×64. The text encoder is from OpenCLIP ViT-H/14 [15]. The training data is WebLI [14], an image-language dataset built from public web images and texts. We use v-prediction as the objective [68], a batch size of 2048, and train for 3M steps. For reference, this text-to-diffusion model achieves an FID of 5.37 on 30K 64×64 MS-COCO 2014 validation images.

**Image backbone and attentional pooler.** We utilize a pre-trained CoCa ViT-L model with input resolution 288×288 as the image backbone, and freeze it during De-Diffusion training [20, 85]. This CoCa model is pre-trained on JFT-3B [72] and ALIGN datasets [42]. Our attentional pooler is equipped with 75 queries, in addition to the `[SOS]` and `[EOS]` tokens to fully utilize the 77 context length defined by CLIP text encoder [15, 59]. The attention pooler has five Transformer blocks which are always randomly initialized.

**Training of De-Diffusion.** The De-Diffusion training data also comes from WebLI [14], while only the images but not the text are used. The broad domain coverage of WebLI enables zero-shot and few-shot evaluations of De-Diffusion on downstream applications in the next section (Sec. 4). For memory efficiency, we use the Adafactor optimizer [70] and a weight decay ratio of 0.01. We train with a batch size of 2048 for 500K steps. The learning rate starts at 3e-4 and is annealed to 3e-6 with cosine decay [52], along with a 10K step warmup [26]. The Gumbel-softmax temperature begins from 2.0 and is exponentially annealed to 0.3 through the entire schedule, which we find is sufficient.

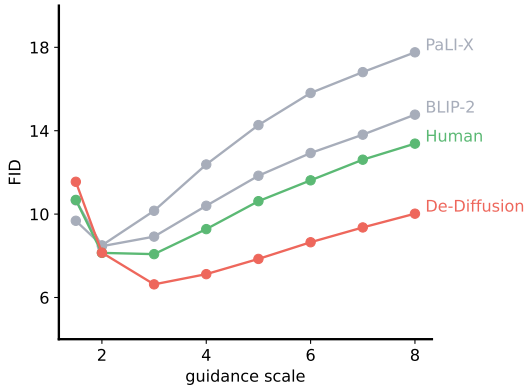


Figure 3. **Evaluating different captioning methods by text-to-image reconstruction.** The text-to-image model is a Stable Diffusion v2-base model [64]. We report FID (↓) on 30K MS-COCO (2014) validation split with 256×256 images. De-Diffusion obtains better FID than human-annotated COCO captions, BLIP-2 [46] (fine-tuned on MS-COCO), and PaLI-X [12] (a multi-task captioning model). Numerical results are provided in the Supp.

## 4. Experiments and Applications

In this section, we introduce several applications of De-Diffusion text, ranging from transferable prompts for text-to-image tools and few-shot vision-language understanding. To demonstrate the versatility of De-Diffusion text across different tasks and domains – that is, its ability to serve as a strong cross-modal interface – all the applications use text from *a single* De-Diffusion model detailed in Sec. 3.2.

### 4.1. Transferable Text-to-Image Prompt

Since De-Diffusion encodes an input image into text and decode it by a text-to-image diffusion model, it is trivial for De-Diffusion text to serve as a prompt suggestion to reconstruct an image by this specific text-to-image diffusion decoder. Furthermore, we demonstrate that De-Diffusion text is transferable to other unseen decoders, *i.e.*, text-to-image tools, such as Imagen [67], Stable Diffusion [64] and Midjourney [3]. This suggests that De-Diffusion text is not over-fitted to a single text-to-image decoder but generalizable across different text-to-image frameworks, which is crucial to make a cross-model interface.

We quantitatively evaluate the ability of De-Diffusion text to transfer to other text-to-image diffusion models and compare with traditional captioning methods. To do this, we develop a benchmark that uses a third-party pre-trained text-to-image model to reconstruct an image from either De-Diffusion text or captions. Specifically, we first obtain De-Diffusion text and captions for a given image. Both are then input into the third-party text-to-image model to synthesize the corresponding image. We compare the synthesized image to the original. Text containing more precise and comprehensive descriptions allows the model to produce images more similar to the original. By evaluating the similarity between the original and the synthesized, we quantify the precision and comprehensiveness of methods.

We use the pre-trained Stable Diffusion v2-base [64] as a generic text-to-image generator, whose weights and training data are oblivious to both De-Diffusion and captioning methods. We measure the similarity between original and synthesized 256×256 images using FID (Frechet Inception Distance) [30] on 30K images from MS-COCO 2014 validation split [13]. Image generation utilizes different classifier-free guidance [35] scales from 1.5 to 8.0, along with 50 steps of DDIM sampling [71].

We evaluate De-Diffusion, human captions and two state-of-the-art image captioning methods, plotted in Fig. 3:

(*i*) Human-annotated captions from MS-COCO provide a strong FID baseline of 8.08 at guidance scale 3.0. We synthesize new images using the longest of the five annotated captions, which we find works best. Other options to utilize human captions are discussed in the Supp.

(*ii*) BLIP-2 refers to its ViT-g OPT 2.7B variant [46], which is fine-tuned on MS-COCO. As one of the state-of-the-art captioning methods, BLIP-2's FID curve is close to that of human-annotated captions.

(*iii*) PaLI-X [12] performs fine-tuning on multiple caption datasets, instead of solely on MS-COCO. As a result, its FID curve is higher than that of BLIP-2.

(*iv*) De-Diffusion is trained with solely web images, but not MS-COCO images or any human-annotated captioning data. It has an indirect access to noisy web image-language pairs through the pre-trained diffusion model. However, De-Diffusion achieves the lowest FID of 6.43 at guidance 3.0, significantly better than the human-annotated captions.

These results indicate that De-Diffusion text precisely and comprehensively verbalizes image details, allowing it to effectively transfer to other text-to-image tools. We provide more qualitative results in the Supp.

### 4.2. Multi-Modal Few-Shot Learner

We next show that De-Diffusion can convert an off-the-shelf LLM, which is never trained on vision-language data, to perform open-ended vision-language task by simply prompting the LLM with few-shot examples, and no adaptive training is required.

LLMs exhibit surprising generalization ability with few-shot learning, adapting to new tasks from just a few annotated task-specific examples without any further training [11]. However, these powerful models are limited to text. Since then, methods have emerged to enable multi-modal capabilities by encoding images into the word embedding space [59, 74] or training a new module to connect vision and language embeddings [4, 46]. However, these approaches have downsides – not only would they introduce prohibitively heavy computational costs due to joint training with enormous language models like 540B PaLM [17],

| methods | LLM | trainable params. | shot | VQAv2 test-dev | OKVQA val | COCO test |
|---|---|---|---|---|---|---|
| BLIP-2 ViT-g [46] | FlanT5$_{XXL}$ | 108M | 0 | 65.0$^{\dagger}$ | 45.9$^{\dagger}$ | - |
| LENS [9] | FlanT5$_{XXL}$ | 0 | 0 | 62.6 | 43.3 | - |
| AnyMAL ViT-G [55] | Llama2$_{70B}$ | - | 0 | 64.2 | 42.6 | 95.9 |
| PICa-Full [84] | GPT-3 | 0 | 16 | 56.1 | 48.0 | - |
| IDEFICS-80B [45] | Llama$_{65B}$ | 14B | 0 | 60.0 | 45.2 | 91.8 |
| IDEFICS-80B [45] | Llama$_{65B}$ | 14B | 4 | 63.6 | 52.4 | 110.3 |
| IDEFICS-80B [45] | Llama$_{65B}$ | 14B | 32 | 65.9 | 57.8 | **116.6** |
| Flamingo-80B [4] | Chinchilla$_{70B}$ | 10B | 0 | 56.3 | 50.6 | 84.3 |
| Flamingo-80B [4] | Chinchilla$_{70B}$ | 10B | 4 | 63.1 | 57.4 | 103.2 |
| Flamingo-80B [4] | Chinchilla$_{70B}$ | 10B | 32 | 67.6 | 57.8 | <u>113.8</u> |
| De-Diffusion ViT-L | PaLM 2-S | 135M | 0 | 63.9 | 51.4 | 63.4 |
| De-Diffusion ViT-L | PaLM 2-S | 135M | 4 | 64.0 | 53.5 | 87.1 |
| De-Diffusion ViT-L | PaLM 2-S | 135M | 32 | 63.1 | 53.3 | 92.0 |
| De-Diffusion ViT-L | PaLM 2-L | 135M | 0 | 67.2 | 57.0 | 88.5 |
| De-Diffusion ViT-L | PaLM 2-L | 135M | 4 | <u>67.9</u> | <u>58.2</u> | 100.3 |
| De-Diffusion ViT-L | PaLM 2-L | 135M | 32 | **68.4** | **60.6** | 103.7 |

Table 1. **Vision-language few-shot learning.** We report VQA accuracy [6] for visual question answering on VQAv2 [27] and OKVQA [54] in the open-ended setting, and CIDEr [77] for MS-COCO image captioning [13]. The **Bold** denotes the top performance and the <u>underlined</u> denotes the second-best in each column. $^{\dagger}$ in-domain COCO images are used for training.

| methods | VQAv2 0-shot | VQAv2 4-shot | VQAv2 32-shot | OKVQA 0-shot | OKVQA 4-shot | OKVQA 32-shot |
|---|---|---|---|---|---|---|
| BLIP-2 OPT$_{2.7b}$ caption [46] | 63.1 | 63.0 | 62.8 | 58.5 | 57.6 | 59.1 |
| Human caption [13] | 63.1 | 63.2 | 63.6 | **59.0** | **58.9** | 60.1 |
| De-Diffusion ViT-L | **65.2** | **66.0** | **66.2** | 57.0 | 58.2 | **60.6** |

Table 2. **Compare to other captions** on the val split of VQAv2 and OKVQA. BLIP-2 represents the top captioning model. Human captions are from MS-COCO annotations. PaLM 2-L is used.

but the visual embeddings also bind to a specific language model such that changing the language model requires retraining. This limits the flexibility of these multi-modal models to keep pace with rapid progress in LLMs.

Unlike previous methods based on deep embeddings, De-Diffusion encodes images into text that any language model can readily comprehend. This allows off-the-shelf language models to ground images by simply interleaving task instructions and De-Diffusion text in any order, as Fig. 1 shows. Using text as a cross-modal interface, De-Diffusion empowers off-the-shelf language models with multi-modal abilities. We next demonstrate that this modular approach achieves state-of-the-art performance on different multi-modal few-shot learning benchmarks, thanks to the comprehensive image context provided by De-Diffusion text, and seamless integration with advanced reasoning abilities provided by the LLMs.

**Multi-modal few-shot learning.** We follow the evaluation protocol of Flamingo [4] to assess few-shot learning on three vision-language tasks including VQAv2 [27], OKVQA [54] and MS-COCO caption [13]. De-Diffusion text for the support images is interleaved along with their questions, answers, and captions to form prompts for the LLMs. The LLM's completion is considered a correct answer only if it exactly matches the ground truth. More details are in the Supp. Results are shown in Tab. 1.

Thanks to the modular nature of De-Diffusion text, we are able to couple the same set of De-Diffusion text with different language models, PaLM 2-S and PaLM 2-L [5] without multi-modal training. The performance of De-Diffusion text paired with PaLM 2-L increases from zero-shot to 32-shot setup on all three tasks. However, when coupled with PaLM 2-S, the 32-shot performance slightly decreases on two VQA benchmarks compared to using four shots. We hypothesize this is because smaller language models like PaLM 2-S benefit less from long context [81], *e.g.*, the around 3600-token prompts for 32 shots.

De-Diffusion text paired with PaLM 2-L matches other methods on MS-COCO captioning, and establishes new state-of-the-art results on two VQA benchmarks for all zero-shot, 4-shot, and 32-shot settings. Meanwhile, De-Diffusion training is also more lightweight in both data and computation. Data-wise, De-Diffusion only uses images, unlike Flamingo and its followups [4, 8, 45] using massive interleaved web text and images, or BLIP-2 [46] which needs human annotations. Computation-wise, De-Diffusion not only uses far fewer parameters (135M in De-Diffusion *vs.* 10B in Flamingo-80B), but its training also does not involve inference with frozen LLMs like 70B-parameter Chinchilla [36] in Flamingo. Instead, it only requires frozen 600M U-Net and CLIP text encoder (Sec. 3.2).

Our results suggest that LLMs, without any multi-modal training, can make grounded inferences for vision-language tasks using just text descriptions of images. The benefits of language models are more pronounced on challenging situations requiring reasoning and commonsense knowledge, such as Outside Knowledge VQA (OKVQA) [54]. As the examples in Fig. 4 show, LLMs can answer non-trivial visual questions that demand both De-Diffusion image context and commonsense knowledge.

On the other hand, the results suggest that De-Diffusion text comprehensively captures semantic elements in images, which language models can leverage to perform VQA and captioning. In Tab. 2, De-Diffusion text outperforms the state-of-the-art captioning method BLIP-2 [46] and even human-annotated captions on VQAv2 [27]. It also surpasses both on OKVQA [54] given 32 shots. This suggests the strength of De-Diffusion text in providing rich context.

**Open-ended one-shot classification.** We follow the protocol from Frozen [74] to evaluate open-ended one-shot image classification on miniImageNet [79]. We interleave De-Diffusion text for the support images along with their real

reference image

[**LLM prompt**] Answer the question given the context.
Image context: a colcandidenverlanticcloseup former recent train train parked traditionenclosed metrotram in a red livery it on railroad platform containing wearing a a yellowpolsurround a knob beside platform near a under platform shelter right there and roof shadows and platform and tracks etc wore worn worn maumaroon brown white stripes markings contentworn yellow yellow stripes train pretoria namibia railway platform train operator worn brown windows platform platform
Image question: What other big vehicle is often painted about the same shade as this vehicle? Short answer:

[**LLM completion**] fire truck.

[**GT answers**] firetruck, fire truck

Figure 4. **VQA with an off-the-shelf LLM**, where De-Diffusion text of the reference image is inserted after "Image context" in the LLM prompt. The LLM then completes the prompt to answer the visual question. De-Diffusion text provides abundant visual details, *e.g.*, *red livery* of the *train*. We use PaLM 2-L [5] as the LLM. Samples are from OKVQA [54].

reference image

[**Chatbot prompt**] Image context: an landsapiccinemageneric photograph dog labrador aus white creamy labrador retriever lying lying resting threshold lying an onto slate footpath pathway street milan ositalian retristreet stil relating called an cream dog shown sleeping sleeping beside near an blue left bicycle bicycle left crowded street left tyre and umbrella blurry beige brown monochrome left left towards and sitting among people street gray walls alley mostly brown buildings street blur street pathway street.
What is amusing about this image?

[**ChatGPT-3.5**] The amusing aspect of the image seems to be the juxtaposition and contrast between the serene and peaceful scene of the Labrador retriever peacefully resting on the slate footpath, and the busy, crowded, and somewhat chaotic background of a bustling street in Milan.

[**Bard**] The amusing thing about the image is the juxtaposition of the sleeping dog and the busy street. The dog is completely relaxed and oblivious to the hustle and bustle around it. The contrast between the two creates a sense of humor.

Figure 5. **Multi-modal dialogue with off-the-shelf text-only chatbots**, where De-Diffusion text is inserted after "Image context" in the text prompt for ChatGPT-3.5 and Bard.

| methods | LLM | w/o induction | w/ induction |
|---|---|---|---|
| P>M>F [39] | - | 95.3 | |
| Frozen [74] | Frozen [74] | 0.9 | 33.8 |
| LQAE [51] | GPT3.5 [11] | 1.0 | 45.9 |
| $\text{SPAE}_{\text{PaLM}}$ [87] | PaLM 2-L [5] | 23.6 | 67.0 |
| De-Diffusion | $\text{Llama2}_{\text{70B}}$ [73] | 64.8 | 87.9 |
| De-Diffusion | PaLM 2-S [5] | 66.4 | 88.6 |
| De-Diffusion | PaLM 2-L [5] | **71.8** | **97.0** |

Table 3. **Open-ended 5-way 1-shot cls. on miniImageNet**, where only the exact class names predicted by the LLM are considered correct. Task induction is introductory text explaining the classification task and providing expected class names at the start of the prompt. Previous best in the closed form is de-emphasized.

class names as prompts for the LLM. The text generated by the LLM is used as the prediction.

We evaluate in an open-ended fashion, where only generating the exact class name is considered correct. There is also an option of task induction, which is introductory text explaining the classification task and providing expected class names at the beginning of the prompt, *e.g.*, "`Classify the context into dog or cat.`" More details are in the Supp.

The results are shown in Tab. 3. Task induction largely increases performance because it helps the language model to generate the exact class names required for open-ended evaluation. With three different LLMs, LLaMA-70B [73], PaLM 2-S and PaLM 2-L [5], De-Diffusion significantly outperforms previous methods. PaLM 2-L inference with task induction achieves 97.0% accuracy, even surpassing the previous closed-form state-of-the-art of 95.3% systematically. These results suggest De-Diffusion excels at verbalizing class names of main objects in images.

### 4.3. Multi-Modal Dialogue

Chatbots such as ChatGPT-3.5 [57] and Bard [1] are LLM-based models that engage users with conversational interactions. They have demonstrated impressive advances in natural language understanding, generation, and conversational capabilities. These chatbots can engage in remarkably human-like dialogue, answer follow-up questions, and perform helpful tasks. However, as language models, they lack grounding in the visual world. In Fig. 5, we demonstrate that De-Diffusion text can provide this missing visual grounding. By incorporating De-Diffusion descriptions of images into the conversational context, chatbots can leverage the rich visual details captured in the text. This allows them to answer challenging questions that require complex reasoning and commonsense knowledge. Furthermore, we find De-Diffusion text transfers across different chatbots.

| tokens | FID↓ | acc. |
|---|---|---|
| 5 | 9.19 | **97.8** |
| 15 | 7.42 | 97.6 |
| 45 | 6.95 | 97.0 |
| 75 | **6.43** | 97.0 |

(a) **Number of tokens**.

| punctuation | FID↓ | acc. |
|---|---|---|
| ✓ | 6.85 | 96.8 |
| × | **6.43** | **97.0** |

(b) **Excluding punctuation**.

| blocks | FID↓ | acc. |
|---|---|---|
| 3 | 6.85 | 96.6 |
| 5 | **6.43** | **97.0** |
| 9 | 6.76 | 93.1 |

(c) **Pooler depth**.

| arch. | init. | # steps | FID↓ | acc. |
|---|---|---|---|---|
| ViT-Base | CoCa | 300K | 6.84 | 92.6 |
| ViT-Large | CoCa | 300K | **6.43** | **97.0** |
| ViT-Large | rand | 300K | 14.6 | 67.2 |
| ViT-Large | rand | 500K | 11.0 | 72.2 |

(d) **Image backbone**.

Table 4. **De-Diffusion ablation experiments.** We evaluate text-to-image reconstruction FID (↓) on MS-COCO (2014) validation split using 256×256 images with Stable Diffusion v2-base. We report the best FID across guidance scales. We also report open-ended 5-way 1-shot classification accuracy on miniImageNet. Default settings are marked in gray.

### 4.4. Ablation

In this section, we ablate different design choices of De-Diffusion. By default, the encoder is a frozen CoCa pre-trained ViT-Large model, and we train De-Diffusion for 300K steps. For text-to-image reconstruction, we use FID on Stable Diffusion v2.0-base, the same setting as Fig. 3, reporting the lowest FID across guidance scales. For few-shot learning, we use 5-way 1-shot classification accuracy on miniImageNet with task induction, identical to Tab. 3.

**Number of tokens.** De-Diffusion text by default uses up all 75 tokens from the CLIP text encoder context. In Tab. 4a, we show performance using 5, 15, and 45 tokens. With more tokens, reconstruction with Stable Diffusion improves, with FID decreasing from 9.19 to 6.43. This aligns with our intuition that longer text descriptions as prompts lead to better text-to-image reconstruction. Interestingly, few-shot classification accuracy decreases from 97.8% to 97.0% with longer text. This suggests when context length is limited, De-Diffusion prioritizes the most salient semantic concepts, usually the image classes. This aligns with the training objective of De-Diffusion to find the most representative text latent to minimize reconstruction error of autoencoding. With longer context, De-Diffusion text includes more comprehensive but subtle concepts beyond the classes, important for reconstruction but not classification.

**Excluding punctuation.** We use the 49K token vocabulary of CLIP as the codebook of latent representations. This naturally results from using the CLIP text encoder for the text-to-image diffusion model. However, we exclude punctuation from the vocabulary, which accounts for around 6% of the original tokens. By excluding these, we can devote more of the limited 75 latent tokens to content words, allowing more semantic concepts to be expressed. In Tab. 4b, we vary these choices. Excluding punctuation improves reconstruction FID on Stable Diffusion from 6.85 to 6.43, suggesting better transferability of De-Diffusion text to other text-to-image models, likely due to the use of more content words. On the other hand, few-shot accuracy on miniImageNet only drops 0.2%, showing punctuation has a small influence on few-shot learning ability when using LLMs.

**Pooler depth.** Tab. 4c varies the depth, *i.e.*, number of Transformer blocks, in the attentional pooler of the image-to-text encoder. Too few layers may limit its ability to capture all the necessary semantics. But too many layers could overfit to the specific text-to-image diffusion model and hurt generalizability. Experiments suggest that with as few as three Transformer blocks, the attentional pooler can effectively transform image features from the pre-trained CoCa backbone into De-Diffusion text. With five blocks, we obtain the best performance on both reconstruction FID with Stable Diffusion and few-shot accuracy on miniImageNet.

**Image backbone.** Tab. 4d varies different image backbone architectures. Increasing the frozen pre-trained CoCa backbone size from ViT-Base to ViT-Large largely improves performance, reducing reconstruction FID from 6.84 to 6.43, and improving few-shot accuracy from 92.6% to 97.0%. We also explore a randomly initialized backbone optimized by the De-Diffusion objective. With 300K training steps, this obtains an FID of 14.6 and few-shot accuracy of 67.2%. Performance increases with a longer 500K schedule. Though still behind pre-trained CoCa backbones, training from scratch achieves 72.2% few-shot accuracy on miniImageNet, surpassing prior methods like SPAE with PaLM 2-L at 67.0%.

## 5. Conclusion

We propose De-Diffusion, an autoencoder whose latent is text representation. By employing a pre-trained text-to-image diffusion model as the decoder, we obtain content-preserving and semantically meaningful textual descriptions for the input images. We then apply De-Diffusion text into text-to-image reconstruction, where De-Diffusion text surpasses human-annotated captions, and combine with advanced LLMs to perform multi-modal few-shot learning, where we surpass large-scale vision-language models. Our results suggest that text representation, like how it connects human perception and cognition, can serve as a strong cross-modal interface for multi-modal tasks.

**Acknowledgement** We thank Jason Baldridge, Nanxin Chen and Yonghui Wu for valuable feedback and support. CW and AY are supported by ONR N00014-23-1-2641.

## References


[1] Bard. hhttps://bard.google.com/chat/. 1, 7

[2] Google I/O 2023: Making AI more helpful for everyone. https://blog.google/technology/ai/google-io-2023-keynote-sundar-pichai. 2

[3] Midjourney. https://www.midjourney.com/home/. 5

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 2, 3, 5, 6

[5] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 1, 2, 6, 7

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 2, 6

[7] Bishnu S Atal and Suzanne L Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The journal of the acoustical society of America*, 1971. 2

[8] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 6

[9] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*, 2023. 3, 6

[10] Thomas G Bever and David Poeppel. Analysis by synthesis: a (re-) emerging program of research for language and vision. *Biolinguistics*, 2010. 2

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 3, 5, 7

[12] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI-X: on scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 5

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 5, 6

[14] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 3, 4

[15] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 4

[16] Noam Chomsky. *Language and mind*. Cambridge University Press, 2006. 2

[17] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 5

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4

[21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[22] Sanja Fidler, Abhishek Sharma, and Raquel Urtasun. A sentence is worth a thousand pixels. In *CVPR*, 2013. 3

[23] Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975. 2

[24] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. *NeurIPS*, 2021. 3

[25] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[26] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 4

[27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, 2017. 6

[28] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 2, 5

[31] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 2006. 2

[32] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006. 2, 3

[33] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NeurIPS*, 1993. 3

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 4

[35] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[36] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *NeurIPs*, 2022. 3, 6

[37] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 3

[38] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *CVPR*, 2022. 4

[39] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *CVPR*, 2022. 7

[40] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*, 2021. 3

[41] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 4

[42] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3, 4

[43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 2

[45] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang,

Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELISC: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023. 3, 6

[46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 5, 6

[47] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 3

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2

[49] Anthony Z Liu, Lajanugen Logeswaran, Sungryull Sohn, and Honglak Lee. A picture is worth a thousand words: Language models plan from pixels. *arXiv preprint arXiv:2303.09031*, 2023. 3

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3

[51] Hao Liu, Wilson Yan, and Pieter Abbeel. Language quantized autoencoders: Towards unsupervised text-image alignment. *arXiv preprint arXiv:2302.00902*, 2023. 3, 7

[52] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 4

[53] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017. 4

[54] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 2019. 6, 7

[55] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. AnyMAL: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*, 2023. 3, 6

[56] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2

[57] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3, 7

[58] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 2023. 3

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4, 5

[60] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 3

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

[62] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 3, 4

[63] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019. 3, 4

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5

[65] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[66] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. *Parallel Distributed Processing*, 1986. 3

[67] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2, 4, 5

[68] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 4

[69] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 3

[70] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, 2018. 4

[71] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[72] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 4

[73] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *CVPR*, 2022. 7

[74] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021. 5, 6, 7

[75] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 3, 4

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 4

[77] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[78] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010. 3

[79] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, 2016. 6

[80] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 3

[81] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023. 6

[82] Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. Neural scene de-rendering. In *CVPR*, 2017. 3

[83] Lingxi Xie, Xiaopeng Zhang, Longhui Wei, Jianlong Chang, and Qi Tian. What is considered complete for visual recognition? *arXiv preprint arXiv:2105.13978*, 2021. 2

[84] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *AAAI*, 2022. 3, 6

[85] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 3, 4

[86] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2, 3

[87] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A. Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, Kevin Murphy, Alexander G. Hauptmann, and Lu Jiang. SPAE: Semantic pyramid autoencoder for multimodal generation with frozen llms. *arXiv preprint arXiv:2306.17842*, 2023. 3, 7

[88] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 3

[89] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 2006. 2

[90] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 3