Optional semester project report – text-to-image generation for multimodal large language models

Tran Minh Son Le
EPFL
son.le@epfl.ch

Abstract

Recent advances in multimodal large language models (MLLMs) deliver good performance on multimodal tasks, e.g, visual question answering, image/video captioning, retrieval, etc. At the same time, there has been a lot of progress in text-to-image generation thanks to the advancements in diffusion modeling, leading to many creative downstream use cases. To achieve such success, these diffusion models are trained on internet-scale data, and they have learned a lot of knowledge about the visual world and how visual concepts are usually described in natural language. In this project, we explore the possibility of obtain a MLLM that exploits the intrinsic knowledge of such diffusion models. For this purpose, we instantiate two architectures, with the first one using synthesized images to help MLLMs solve text-only tasks and the second one (still work in progress) bypassing the image generation step and directly enable a LLM to understand feature representations learned by Stable Diffusion. By conducting extensive experiments and ablation studies, despite some drawbacks, we obtain promising results, showing the potential of diffusion models as visual interpreters in the MLLM framework, especially when input images are not available.

1 Introduction

Recent advances in multimodal large language models (MLLMs) deliver good performance on multimodal tasks, e.g, visual question answering, image/video captioning, retrieval, etc. A notable example is GPT4-Vision (Achiam et al., 2023). At the same time, there has been a lot of progress in text-to-image generation thanks to the advancements in diffusion modeling, leading to many creative downstream use cases. A notable (closed-source) example is DALL·E 3 (Betker et al., 2023). To achieve such success, these diffusion models are trained on internet-scale data, and they have learned a lot of knowledge about the visual world and how visual concepts are usually described in natural language. In this project, we explore the possibility of obtain a MLLM that exploits the intrinsic knowledge of such diffusion models.

Concretely, we propose two architectures:

- Architecture 1 (Fig. 1a): Use an off-the-shelf diffusion model to generate an image from the input text-only task prompt, then feed both into an off-the-shelf MLLM to predict the text output.
- Architecture 2 (Fig. 1b): Use the diffusion model, e.g., Stable Diffusion (Rombach et al., 2022) as a visual encoder of the MLLM, bypassing the image generation. At the moment, this is work in progress, and we only train it as a regular MLLM that requires both image and text inputs to reason visually.

By conducting extensive experiments and ablation studies, despite some drawbacks, we obtain promising results, showing the potential of diffusion models as visual interpreters in the MLLM framework, especially when input images are not available.

2 Related work

Multimodal large language models. In the current era of large language models, there has been a recent resurgence of vision-language models (VLMs), especially image-to-text models. Different from a previous iteration of VLMs based on encoder-only language models (Lu et al., 2019, *inter alia*), the BLIP model series (Li et al., 2022a, 2023b; Dai et al., 2023) introduces the idea of aligning visual features, usually CLIP-based features (Radford et al., 2021), to the text token embedding space of a LLM via a multimodal projector module. This recipe turns out to be successful and thus adopted by many contemporary works (Huang

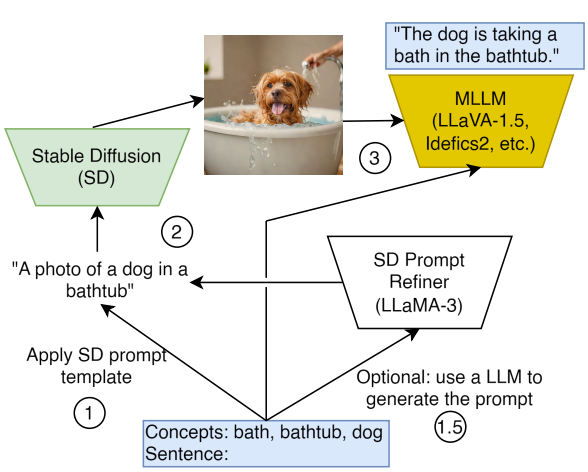
et al., 2023; Peng et al., 2024; Ye et al., 2023, 2024; Zhu et al., 2024; Liu et al., 2023b, 2024a). However, the main limitation of these methods is that they do not support more complex image-text configurations such as interleaved image-text sequences or multi-image inputs. The key idea to tackle this challenge is the inclusion of interleaved web-scale image-text data into the training data mix (Alayrac et al., 2022; Awadalla et al., 2023; Li et al., 2023a; Zhao et al., 2024; Laurençon et al., 2023, 2024; Sun et al., 2024a,b; Lin et al., 2024). These methods demonstrate that such a training strategy unlocks the MLLM's ability to perform multimodal in-context learning. Please refer to surveys (Wadekar et al., 2024; Caffagni et al., 2024; Bai et al., 2024; Yin et al., 2023; Bordes et al., 2024; Zhang et al., 2024a; Song et al., 2023) as well as ablation studies (Karamcheti et al., 2024; McKinzie et al., 2024; Lin et al., 2024) for a thorough understanding of this rapidly evolving field.

Visually-augmented natural language modeling. Similar to VLMs, one line of research focuses on enhancing language models with either generated or retrieved visual inputs, hoping to boost their performance on tasks such as natural language understanding (NLU), natural language generation (NLG), or machine translation (MT). iACE (Lu et al., 2022) trains a cross-modal encoder by distilling from pre-trained vision-language backbones, where images are generated by a VQGAN (Esser et al., 2021; Crowson et al., 2022). Meanwhile, Z-LaVI (Yang et al., 2022) implements a zero-shot pipeline by enhancing language models' predicted probabilities with CLIP-based (Radford et al., 2021) image-text similarity scores, where images are both retrieved via a search engine and generated by DALL·E (Ramesh et al., 2021). VaLM (Wang et al., 2023) and MORE (Cui et al., 2024) retrieve images, while LIVE (Tang et al., 2023b) uses Stable Diffusion (Rombach et al., 2022) to generate them. Both works fuse visual features with textual features deep within the language model decoder via cross-attention. Meanwhile, iNLG (Zhu et al., 2023) has a similar architecture to the MLLMs discussed above and introduces a contrastive loss to enforce similarity of the generated text to the visual features. Arguably, one drawback of these methods is the added time to generate or retrieve images. Thus, VAWI (Guo et al., 2023a) bypasses this step by directly using the CLIP (Radford et al., 2021) text features as a proxy to visual features. Similarly, Hagström and Johansson (2022) proposes different methods to adapt a VLM to a text-only input. Notably, all these methods use CLIP as the visual encoder. Meanwhile, in machine translation, visual inputs (either given or generated) are employed as extra context to enhance translation quality (Li et al., 2022b; Guo et al., 2023b, *inter alia*). Different from these methods, which all train from scratch the model's ability to understand images or side information, our pipeline utilizes pre-trained MLLMs with strong zero-shot capabilities, augmented by images synthesized by state-of-the-art text-to-image diffusion models (Podell et al., 2024; Sauer et al., 2023; Chen et al., 2024).

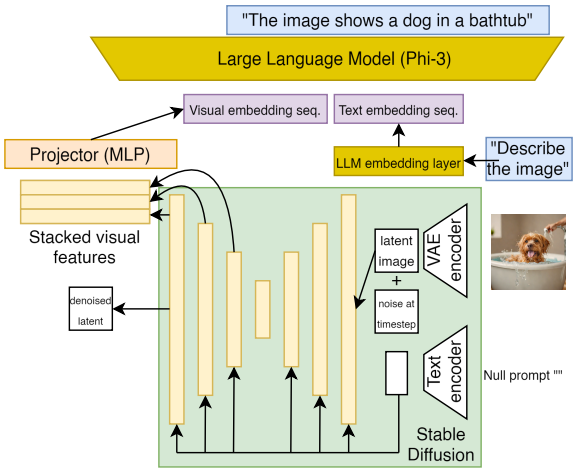
Using diffusion models for downstream tasks. Recent text-to-image diffusion models have been popular, notably open-source models such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2024) which is based on the U-Net (Ronneberger et al., 2015) and Pixart-Σ (Chen et al., 2024) which is based on the Diffusion Transformer (Peebles and Xie, 2023). These models, with their rich internal representations, have boosted the usefulness of downstream applications such as image editing (Mokady et al., 2023; Luo et al., 2024, *inter alia*), image-to-image variation (Ramesh et al., 2022), as well as computer vision tasks such as segmentation (Xu et al., 2023; Zhao et al., 2023, *inter alia*) and semantic correspondence (Zhang et al., 2023; Tang et al., 2023a; Luo et al., 2023). Moreover, diffusion models have also been used as a "visual decoder" in MLLMs so that they can generate images with better prompt adherence (Sun et al., 2024b,a; Dong et al., 2024; Koh et al., 2024; Ge et al., 2024). On the other hand, one work (He et al., 2024) attempts to use Stable Diffusion (Rombach et al., 2022), more specifically its VAE (Esser et al., 2021), as the visual encoder. Inspired by these works, we propose a new model that uses the diffusion-based U-Net as the visual encoder in the LLaVA (Liu et al., 2023b) architecture.

3 Methodology

Problem statement. In this project, we examine whether modern LLMs, e.g., LLaMA-2 (Touvron et al., 2023), are able to perform better on text-only tasks such as multiple-choice question answering (MCQA) and natural language generation (NLG) when they are incorporated into a MLLM framework, e.g., LLaVA (Liu et al., 2023b). For this purpose, inspired by visually-augmented language



(a) Architecture 1: Our proposed pipeline for enhancing MLLMs on text-only tasks with synthetic image generation. In step 1, we extract a prompt for the diffusion model to generate an image (step 2), after which we feed both the synthesized image and the input task prompt into the MLLM (step 2).



(b) Archicture 2: Our proposed method to equip a LLM, Phi-3 (Abdin et al., 2024), with a T2I diffusion-based visual encoder, Stable Diffusion (Rombach et al., 2022). Our current MLLM architecture closely follows (Liu et al., 2024a) and (Karamcheti et al., 2024).

Figure 1: Our two instantiations of incorporating text-to-image diffusion models into the MLLM framework. The example input-output pair in both figures (in blue) is taken from the CommonGen dataset (Lin et al., 2020). The picture was generated using SDXL-Turbo (Sauer et al., 2023).

models (Tang et al., 2023b; Zhu et al., 2023), we investigate using MLLMs whose inputs are the input texts of the given textual task and synthetic images generated by diffusion-based models when prompted with the input text (e.g., the question of a MCQA sample).

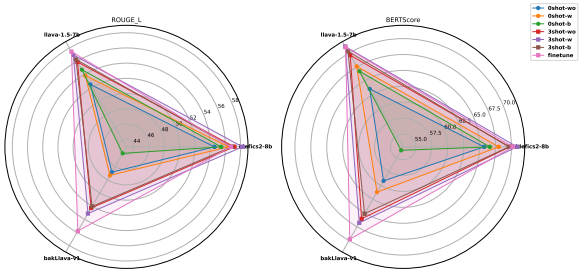
3.1 Preliminary: Multimodal LLMs

Contemporary MLLM architectures usually have three modules: a visual encoder, a multimodal projector, and a LLM decoder. The idea is to enable pre-trained LLMs to understand visual inputs such as images and videos encoded by pre-trained visual foundation models such as CLIP (Radford et al., 2021) or DINO (Oquab et al., 2024). The projector then projects visual features into the LLM space as embeddings of "visual tokens", to be prepended to the text token embeddings. This whole concatenated sequence is then fed into the LLM, which generates new text tokens autoregressively. The learning objective is usually the regular language modeling objective, i.e., the autoregressive factorization of the joint log-likelihood of the input sequence, conditioned on the visual embeddings.

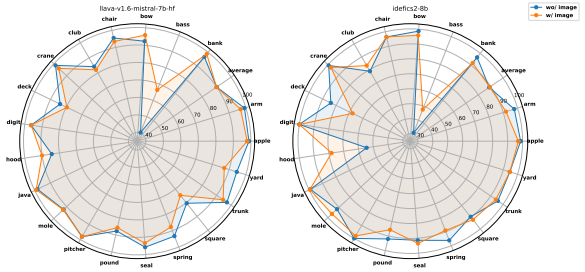
There are usually two stages when training a MLLM. In the first stage, called the pre-training stage or alignment stage, the MLLM learns to align visual features with text features. Here, only the projector is trainable. This stage often uses paired text-image data, such as captions (e.g., Sharma et al., 2018). In the second stage, called the fine-tuning stage or visual instruction tuning (Liu et al., 2023b), both the projector and the LLM are fine-tuned on multimodal instruction data such as visual question answering (e.g., Goyal et al., 2017).

3.2 Preliminary: Latent diffusion models

A text-to-image (T2I) diffusion model generates an image by iteratively denoising a noisy image over many timesteps. In this project, we only consider diffusion models that operate in a latent space, e.g., Stable Diffusion (Rombach et al., 2022) (SD) instead of the pixel space (e.g., Saharia et al., 2022). A T2I diffusion model has three components: a text encoder that encodes the user-provided text prompt, a VQGAN-based VAE (Esser et al., 2021) that encodes and decodes pixels into a latent space, and a U-Net (Ronneberger et al., 2015) or Diffusion Transformer (DiT) (Peebles and Xie, 2023) to predict the noise applied to the input (latent) image. During inference, the generation process begins with a random vector, which is then iteratively denoised over multiple timesteps, guided by the encoded text embeddings. Such textual conditioning is enabled by cross-attention layers in the U-Net or adaptive layer norm in the DiT (Peebles and Xie, 2023). In this project, we synthesize images with SDXL-Turbo (Sauer et al., 2023), a model distilled from SDXL (Podell et al., 2024) and generates an image in only a few timesteps.



(a) CommonGen results. `wo`: without image, `w`: with image; `b`: we use a constant black image as an ablation; `finetune`: model is fine-tuned on the whole train split.



(b) F1 scores across different words (and the macro-average). Using synthetic images helps for some words but hurts for other words, resulting in an unchanged average score.

Figure 2: Results for CommonGen and CoarseWSD-20. For CommonGen, synthetic images help a lot under the zero-shot setting, whereas they do not help much under the 3-shot setting. On the other hand, fine-tuning does not help much compared to 3-shot prompting; it only teaches the model how to do the task.

3.3 MLLM with synthetic images

As text-only tasks do not provide images in their datasets, we have two intuitive ways to evaluate MLLMs on these tasks. The first approach is to simply discard the visual encoder and the multimodal projector and only use the LLM part of the MLLM to process the textual inputs. This is possible since this module is usually just a fine-tuned version of the base LLM, as discussed above. Alternatively, we can attempt to "imagine" an image from the textual input and feed both into the MLLM.

To do this, we follow iNLG (Zhu et al., 2023) and use diffusion models (Sauer et al., 2023; Chen et al., 2024) to generate images with the prompt being the textual input of the text-only task (e.g., the question of the MCQA or the context of the NLG task). The pipeline, illustrated as Architecture 1 in Fig. 1a, consists of the following three steps:

Step 1: From the input task (e.g., MCQA), we extract a prompt for the T2I diffusion model. This process can be automatic (e.g., use as the prompt the question of a multiple-choice QA sample), or assisted by a LLM (AI@Meta, 2024), where we use a manual few-shot instruction to teach it how to generate/refine such prompts.

Step 2: Feed the extracted prompt to a T2I diffusion model to generate an image.

Step 3: Feed the original input task prompt (e.g., the concatenated question and possible MCQA options) and the synthesized image to the MLLM to generate textual outputs (e.g., the predicted answer).

3.4 MLLM with a SD-based visual encoder

A natural progression of the above approach is to bypass the image generation step and directly use features extracted from the diffusion model. Inspired by He et al. (2024), we replace the CLIP ViT commonly used in most contemporary MLLMs with the U-Net from Stable Diffusion v1.5 (Rombach et al., 2022) (SD1.5). More specifically, we follow Zhang et al. (2023, 2024b) and use the ResNet (He et al., 2016) output features of layers 2, 5, and 8 (out of 12 layers) from the upsampling blocks of the U-Net. Fig. 1b demonstrates our method, Architecture 2. Next, we explain more about our design choices.

In this project, we use the LLaVA-1.5 strategy (Liu et al., 2023b, 2024a), as implemented by Karamcheti et al. (2024). The visual encoder is usually the CLIP Vision Transformer (ViT) (Radford et al., 2021; Dosovitskiy et al., 2021), which embeds images into fixed-resolution visual patches (usually with a resolution of 16^2) during its pretraining. A multimodal projector then projects each patch embedding into the LLM embedding space. The projected visual sequence is then directly prepended to the sequence of text token embeddings and subsequently fed into the LLM for autoregressive sequence modeling. Although a ViT can be modified so that it can input and output at any resolution (Dehghani et al., 2023), increasing the resolution of the input images usually lead to more patches, meaning a longer visual sequence for the LLM. Provided that no resampling mechanism is used, this can boost the performance of the MLLM at the expense of longer training and inference time (Pantazopoulos et al., 2024; Cha et al., 2024) due to the quadratic complexity of the attention mechanism.

Back to SD1.5, its native input/output resolution is 512^2, whereas the U-Net features have differ-

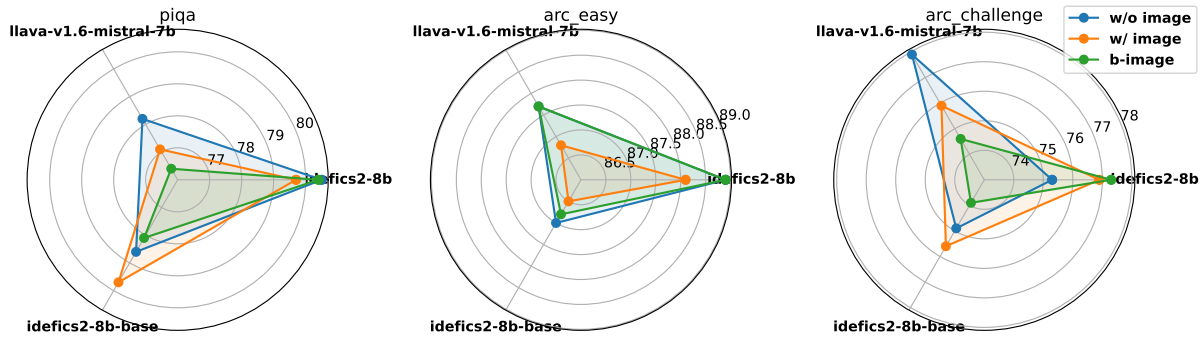


Figure 3: Results on PIQA, AI2 ARC showcasing the impact of using synthetic images. Zero-shot. Black image means we use a constant black image instead of synthesizing it.

ent spatial resolutions (8^2, 16^2, and 32^2 for layers 2, 5, and 8 respectively). In contrast to works that use a projector that resamples the visual sequence to a short length, we simply follow LLaVA-1.5 (Liu et al., 2024a) and use a multilayer perceptron (MLP), which maintains the length of the visual tokens. Thus, at the time of writing this manuscript, we opt to resize each U-Net feature to a fixed size (e.g., 16^2) using naive bilinear interpolation (we leave the method of adaptive average pooling (Pantazopoulos et al., 2024) for future work), then stack them along the channel dimension to create the output visual sequence, which is subsequently fed into the multimodal projector as described above.

In addition, inspired by Karamcheti et al. (2024); Tong et al. (2024); Kar et al. (2024), we explore combining visual features from different visual encoders: SDv1.5 (Rombach et al., 2022), SigLIP (Zhai et al., 2023), and DINOv2 (Oquab et al., 2024). To create an ensemble of features, we currently follow Karamcheti et al. (2024) and simply feed the image to each encoder separately and stack the resulting visual patch representations in the channel dimension. The rest of the MLLM pipeline proceeds as described above.

Using SD as a visual encoder is non-trivial as it was originally trained to denoise images over multiple timesteps, guided by a text prompt (Ho and Salimans, 2021). However, as proof-of-concept, we opt to ablate and simply extract SD features by running only one denoising step with the timestep being 0, meaning there is no added noise, please see Fig. 1b. We also do not feed text prompts into SD and instead use a null prompt (empty string), following the computer vision literature about using diffusion features on real images (Luo et al., 2024; Tian et al., 2024). We leave to future work exploration of more complex diffusion feature extraction strategies: running the denoising process over many timesteps (even starting from random noise – no need for images), ablating over U-Net layer features, using text prompts, or using better diffusion models such as Stable Diffusion 3 (Esser et al., 2024).

4 Experiments

4.1 MLLM with synthetic images

Tasks. We evaluate Architecture 1, MLLM with synthesized images (Fig. 1a), on two text-only tasks [1]. The first task is multiple-choice question answering (MCQA) including the following datasets: PIQA (Bisk et al., 2020), which is about physical commonsense, ARC (Clark et al., 2018), which is about grade-school science, and CoarseWSD-20 (Loureiro et al., 2021), which disambiguates ambiguous words (e.g., apple or Apple) by choosing the correct meaning (sense) in a given context. The second task is text generation with CommonGen (Lin et al., 2020), which generates a sentence given concepts.

Metrics. For MCQA, we use the F1 score to account for class imbalance. For CommonGen, we follow previous work (Wang et al., 2022; Zhu et al., 2023; Tang et al., 2023b; Guo et al., 2023a; Cui et al., 2024) and use BLEU-4 (Papineni et al., 2002) which computes precision in n-gram overlap between predictions and references, ROUGE-L (Lin, 2004) which focuses on recall, METEOR (Banerjee and Lavie, 2005) which combines both precision and recall, CIDEr (Vedantam et al., 2015) which evaluates consensus based on word frequency, SPICE (Anderson et al., 2016) which

[1] Our code is hosted at https://github.com/sonalexle/vlm-viz/

Method	k-shot?	Image?	B-4	R-L	Met.	CIDEr	SPICE	BertS.
MORE-OPT-2.7b (Cui et al., 2024)	FT	✓	32.8	57.1	32.1	17.0	32.9	-
I&V T5-large (Wang et al., 2022)	FT	✓	**40.6**	-	-	17.7	31.3	-
LIVE-T5 (Tang et al., 2023b)	FT	✓	27.9	-	-	15.8	31.4	-
iNLG-T5-large (Zhu et al., 2023)	FT	✓	34.5	-	**33.9**	17.8	**35.5**	72.7
VAWI-SBS-T5-3b (Guo et al., 2023a)	FT	✓	37.5	**59.9**	33.4	**18.3**	34.7	-
llava-1.5-7b (Liu et al., 2024a)	FT	✓	32.5	57.5	33.4	17.1	33.9	71.1
Mistral-7B-v0.1 (Jiang et al., 2023)	0	✗	21.2	49.8	26.6	12.2	25.6	57.4
llava-1.5-7b	0	✗	26.8	52.5	29.8	14.5	28.1	63.5
llava-1.5-7b	0	✓	26.9	53.8	30.4	14.7	29.9	67.6
idefics2-8b-base (Laurençon et al., 2024)	0	✗	30.0	54.3	29.7	15.2	28.4	65.1
idefics2-8b-base	0	✓	31.7	56.2	30.8	15.9	30.4	68.4
gpt-3.5-turbo* (Ouyang et al., 2022)	3	✗	28.9	53.2	31.1	15.9	28.9	-
Mistral-7B-v0.1	3	✗	36.2	58.9	32.3	17.5	32.0	72.5
llava-1.5-7b	3	✗	31.1	55.9	32.1	16.1	30.9	69.6
llava-1.5-7b	3	✓	32.4	57.0	32.6	16.8	32.4	71.2
idefics2-8b-base	3	✗	36.8	59.0	32.7	17.8	33.5	72.6
idefics2-8b-base	3	✓	37.2	59.3	32.9	18.0	33.5	**72.8**

Table 1: CommonGen results. **Bold** is highest, underline is second highest. B-4 is BLEU-4; R-L is ROUGE-L; Met. is METEOR; BertS. is BERTScore. FT means the model was fine-tuned on either the CommonGen train split or some external dataset used in some previous work. ✓ means we feed the generated image into the model, and vice versa for ✗. *: evaluated by Cui et al. (2024).

measures semantic content via scene graphs, and BERTScore (Zhang et al., 2020) uses BERT embeddings (Devlin et al., 2019) to compute semantic (cosine) similarity between words in predicted and reference sentences.

Models. For synthesizing images, we use Stable Diffusion XL (SDXL) Turbo (Sauer et al., 2023), a state-of-the-art open-source diffusion model that generates a high-quality image in just one step. We also experimented with Pixart-Σ (Chen et al., 2024) but the generations were not high-quality. For extracting prompts for T2I, we use LLaMA-3 (Touvron et al., 2023). For MLLMs, we use a variety of models: LLaVA-1.5 (Liu et al., 2024a) (Vicuna-7b (Zheng et al., 2023) backbone), LLaVA-NeXT (Liu et al., 2024b) (Mistral-7b (Jiang et al., 2023) backbone, multi-crop strategy), BakLLaVA [2] (Mistral-7b backbone, LLaVA-1 recipe (Liu et al., 2023b)), and Idefics-2 (Laurençon et al., 2024) (Mistral-7b backbone, interleaved data pre-training, better modeling strategy). Laurençon et al. (2024) provide two checkpoints, Idefics-8b-base and Idefics-8b, where the former is a pre-trained only checkpoint (stage 1 as described in Section 3.1), instead of visual instruction tuning as in the case of the latter (meaning it was trained under both stages 1 and 2).

Image generation strategy. As mentioned above, we use LLaMA-3 (AI@Meta, 2024) to generate SD prompts (SD Prompt Refiner in Fig. 1a), which are then used by SDXL Turbo to generate images. We do these two steps offline before running the benchmarks. For PIQA and ARC, we manually construct a few-shot in-context learning (ICL) instruction prompt for LLaMA-2, where we use examples from the train split as the context. For CommonGen, we use as context the example concepts [3] (the prompt template) given by (Clark et al., 2018). Then, we manually all in-context SD prompt exemplars given each context. Note that this prompt generation step is not necessary; however, the generated images would be of lower quality as current open-source diffusion models' prompt adherence skill is limited (Hu et al., 2024).

Implementation details. Regarding prompt templates, we experimented with using chat templates (i.e., user, assistant roles) and found that not using a template improves the results; thus, we only report results without chat templates. For MCQA, we perform 0-shot inference and format the prompt as "`Question: Options: (A) (B) etc. Answer: (`" (among other strings) and have the model generate a letter (prompt inspired by Li et al. (2024)). For CommonGen, we either fine-tune our models on the full train split, or we conduct 0-shot and

[2] https://github.com/SkunkworksAI/BakLLaVA

[3] https://huggingface.co/datasets/allenai/commongen_lite

Visual encoder	AI2D	GQA	POPE	Tally-QA	Text-VQA	Vizwiz	VQAv2	VSR	Avg.
sd1.5	57.3	58.2	85.4	56.4	12.2	26.6	70.6	61.5	53.5
dino	58.4	61.0	86.3	60.7	12.8	29.1	72.4	59.3	55.0
dinosiglip	61.5	62.3	86.6	62.1	38.1	38.8	76.7	**67.0**	61.6
siglipsd1.5	**61.9**	61.7	86.2	65.3	40.7	39.0	**77.1**	66.4	62.3
siglip	61.8	62.3	85.7	64.9	**41.9**	**41.8**	76.8	63.7	**62.4**
dinosiglipsd1.5	61.2	**62.4**	**87.3**	**66.0**	38.3	39.8	76.9	66.9	**62.4**

Table 2: The LLM decoder is always Phi-3. **Bolded** are best, underlined are second-best. Although the difference between the scores are tiny, we see that combining SD with at least SigLIP provides the best results. Note that results for VQAv2 and TallyQA are obtained from a 16K random subsample of the original splits. All displayed scores are accuracy scores as computed by official evaluation code.

3-shot inference (fixed in-context examples). Under the latter scenario, because these models were instruction-tuned, they often generate extra content besides the requested sentence, hence we apply a post-processing step to automatically extract the first sentence of the generated text. For CommonGen fine-tuning, the train split has roughly 65K samples; we use 1 epoch with a batch size of 32. Whenever we provide images as input, we modify the prompt to explicitly instruct the model to "observe" the image. As an ablation, we follow Hagström and Johansson (2022) and use a constant black image instead of synthesized ones. For all settings where synthetic images are provided to the model, only one image is provided per text input, even when 3-shot prompting is used.

CommonGen results. Results for CommonGen are shown in Fig. 2a and Table 1. Synthesizing images helps especially in the 0-shot case as it gives the model more idea on what kind of sentence to generate – we can think of the desired output as a kind of caption to the synthetic image. On the other hand, when we use three shots, generated images do not seem to help much. This might be due to the fact that the in-context examples might already be sufficient teach the model how to do the task, and thus it does not need to use the synthesized image. On the other hand, using a black image usually hurts the performance, as it might distract the model from actually performing the task, given that the LLM sees the image as a sequence of visual tokens of non-trivial length. Another observation from Table 1 is the difference in 0-shot text-only performance between Idefics2-8b-base and its LLM base model Mistral-7b-v0.1. Note that both models did not undergo any instruction tuning during their construction, although Idefics2's LLM module were LoRA-tuned (Hu et al., 2022) during its pre-training stage (Laurençon et al., 2024). This might suggest a transfer of visual knowledge to the LLM decoder.

CoarseWSD-20 results. CoarseWSD-20 results are shown in Fig. 2b. Here, we observe that synthetic images can help distinguish senses of some words, but it might be a distraction if the word is already easy enough. Thus, the average F1 score is almost unchanged.

PIQA and ARC results. PIQA and ARC results are displayed in Fig. 3. Again, these results not definitive and do not seem to have any pattern, besides the fact that they are model- and data-dependent. One possible explanation is that these tasks, although each sentence of the input prompt might be highly visualizable as the mentioned subjects are usually common nouns, answering the questions do not require visual knowledge and instead requires complex reasoning. This is in contrast to CommonGen and CoarseWSD-20, where the task inputs are usually easier to visualize or "imagine" (concepts for the former, descriptive context for the latter). We leave to future work the quantification of "imageability" of texts (Doostmohammadi and Kuhlmann, 2022; Kastner et al., 2021; Wu and Smith, 2023; Bird et al., 2001). Previous work on visually-augmented NLP has also explored this concept – they generate or retrieve images only when the input is highly visual (Tang et al., 2023b; Guo et al., 2023a), which is estimated by ad-hoc methods such as CLIPScore (Hessel et al., 2021).

4.2 MLLM with SD-based visual encoder

Tasks and metrics. Currently, we use the framework and implementation by Karamcheti et al.

(2024) to train our models. They propose an evaluation suite of multimodal tasks, from which we use AI2D (Kembhavi et al., 2016), GQA (Hudson and Manning, 2019), POPE (Li et al., 2023c), Tally-QA (Acharya et al., 2019), Text-VQA (Singh et al., 2019), Vizwiz (Bigham et al., 2010), VQA-v2 (Goyal et al., 2017), and VSR (Liu et al., 2023a). We also use (Karamcheti et al., 2024)'s implementation of the benchmarks to evaluate our models, where for Tally-QA and VQA-v2 we test on a 16K random subset (all other datasets have around 16K samples or fewer). These tasks are formulated as visual QA, with the official metric being the accuracy score as computed by the official codebases (also provided [4] by Karamcheti et al. (2024)).

Implementation details. As mentioned earlier in Section 3.4, we currently only use the most basic setup to reduce the degree of complexity of our implementation: SDv1.5 U-Net (Rombach et al., 2022) (extracting features from upsampling ResNet output layers 2, 5, 8, empty string, 1-step denoising with no added noise), input resolution 512^2 for SD and 224^2 for DINOv2 (Oquab et al., 2024) and SigLIP (Zhai et al., 2023), visual features always concatenated in the channel dimension (the spatial dimensions all resized to 224^2 using bilinear interpolation), MLP multimodal projector, Phi-3 (Abdin et al., 2024) as the LLM decoder, visual instruction-tuning only (Karamcheti et al., 2024) (the first stage, as mentioned above, is skipped) on the LLaVA-1.5 visual instruction dataset (Liu et al., 2024a). We leave to future work more complex modeling strategies.

Results. The full results can be found in Table 2. Although the SD encoder by itself is not competitive, ensembling it with SigLIP (Zhai et al., 2023) gives promising results. One explanation is that SD features, similar to DINO features, are not aligned with the language space unlike SigLIP features. This is because of their pre-training objectives: denoising diffisuion for SDv1.5, self-supervised learning for DINOv2, and contrastive image-text matching for SigLIP. Despite this naive feature fusion strategy, we observe non-trivial improvements given that each dataset has at least 4K samples. This is thanks to the complementary strengths of each visual encoder (Karamcheti et al., 2024; Tong et al., 2024; Kar et al., 2024): SD is good at spatial and depth understanding (Zhang et al., 2023); DINOv2 is good at discriminative features and semantics (Oquab et al., 2024); and SigLIP is good at vision-text alignment and it understands many concepts (Zhai et al., 2023).

[4] https://github.com/sonalexle/prismatic-vlms

5 Conclusion

In this project, we propose two architectures for MLLMs to exploit the knowledge learned by text-to-image (T2I) diffusion models. In the first architecture, we use T2I generation to directly synthesize images given the input text-only task, and obtained promising results on tasks that are more "visual", although the improvements are small compared to text-only baselines. In the second architecture, we directly replace the visual encoder module in the usual MLLM recipe with the U-Net (Ronneberger et al., 2015) of Stable Diffusion (Rombach et al., 2022). Although the results are not state-of-the-art, we design the architecture with minimal complexity, yet we already observe promising results, suggesting room for improvement.

As future work, we plan to evaluate our architecture 2 on text-only tasks that we evaluated on architecture 1 in this report. We also plan to investigate the concept of imageability (Doostmohammadi and Kuhlmann, 2022; Kastner et al., 2021; Wu and Smith, 2023; Bird et al., 2001), to understand when and if a diffusion model's visual "imaginations" help for texts that are highly "imageable". In addition, we plan to improve architecture 2 itself by exploring more complex modeling decisions, including but not limited to: multiple-step denoising to obtain features from different timesteps (non-trivial to possibly combine them), feeding text prompts to the diffusion-based visual encoder, and do not use images at all and only fine-tune the projector and LLM decoder on diffusion features "imagined" from the text prompt.

Furthermore, we plan to implement more complex multimodal projector architectures (e.g., Kar et al., 2024) to aggregate strengths from diffusion visual branches while maintain high visual fidelity and faithfulness. One idea could be to combine the spatially smaller but semantically stronger features of SigLIP (Zhai et al., 2023) with diffusion's spatially larger but semantically weaker features via feature pyramid network-like architecture (Lin et al., 2017) that aims to imbue larger feature maps with semantics from smaller feature maps. This idea has been used in the diffusion literature (Zhao et al., 2023; Xu et al., 2023).

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card. *Meta LLaMA*.

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

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.

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. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. 2024. A Survey of Multimodal Large Language Model from A Data-centric Perspective. *arXiv preprint arXiv:2405.16640*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *OpenAI*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Helen Bird, Sue Franklin, and David Howard. 2001. Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, 33(1):73–79.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. 2024. An Introduction to Vision-Language Modeling. *arXiv preprint arXiv:2405.17247*.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The (R) Evolution of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2402.12451*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13817–13827.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision (ECCV)*, pages 88–105. Springer.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. More: Multi-modal retrieval augmented generative commonsense reasoning. *arXiv preprint arXiv:2402.13625*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi.

2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Peter Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey A. Gritsenko, Mario Lucic, and Neil Houlsby. 2023. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*.

Ehsan Doostmohammadi and Marco Kuhlmann. 2022. On the effects of video grounding on language models. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 1–6.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning (ICML)*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 12873–12883.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2024. Making LLaMA SEE and draw with SEED tokenizer. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 6904–6913.

Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji-Rong Wen. 2023a. Visually-augmented pretrained language models for NLP tasks without images. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14912–14929, Toronto, Canada. Association for Computational Linguistics.

Wenyu Guo, Qingkai Fang, Dong Yu, and Yang Feng. 2023b. Bridging the gap between synthetic and authentic images for multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2863–2874, Singapore. Association for Computational Linguistics.

Lovisa Hagström and Richard Johansson. 2022. How to adapt pre-trained vision-and-language models to a text-only input? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5582–5596.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 18180–18187.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 6700–6709.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. 2024. BRAVE: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*.

Marc A Kastner, Chihaya Matsuhira, Ichiro Ide, and Shin'ichi Satoh. 2021. A multi-modal dataset for analyzing the imageability of concepts across modalities. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 213–218. IEEE.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*, pages 235–251. Springer.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv preprint arXiv:2305.03726*.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning (ICML)*, pages 12888–12900. PMLR.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. 2022b. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5216–5226.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2125.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Yujie Lu, Wanrong Zhu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4392–4402, Seattle, United States. Association for Computational Linguistics.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. 2024. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. 2023. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems (NeurIPS)*, 35:27730–27744.

Georgios Pantazopoulos, Alessandro Suglia, Oliver Lemon, and Arash Eshghi. 2024. Lost in Space: Probing Fine-grained Spatial Understanding in Vision and Language Resamplers. *arXiv preprint arXiv:2404.13594*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 4195–4205.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning (ICML)*, pages 8821–8831. PMLR.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

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

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 8317–8326.

Shezheng Song, Xiaopeng Li, and Shasha Li. 2023. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.07594*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024a. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14398–14409.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Emu: Generative Pretraining in Multimodality. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023a. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tianyi Tang, Yushuo Chen, Yifan Du, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Learning to imagine: Visually-augmented natural language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9468–9481, Toronto, Canada. Association for Computational Linguistics.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. 2024. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3554–3563.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4566–4575.

Shakti N Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. 2024. The Evolution of Multimodal Model Architectures. *arXiv preprint arXiv:2405.17927*.

PeiFeng Wang, Jonathan Zamora, Junfeng Liu, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. Contextualized scene imagination for generative commonsense reasoning. In *International Conference on Learning Representations*.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Visually-Augmented Language Modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Si Wu and David Smith. 2023. Composition and deformance: Measuring imageability with a text-to-image model. In *Proceedings of the 5th Workshop on Narrative Understanding*, pages 106–117, Toronto, Canada. Association for Computational Linguistics.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-vocabulary panoptic segmentation with text-to-image

diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966.

Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-LaVI: Zero-shot language solver fueled by visual imagination. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1186–1203, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2024b. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3076–3085.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2023. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2023. Visualize before you write: Imagination-guided open-ended text generation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 78–92, Dubrovnik, Croatia. Association for Computational Linguistics.