DREAMLLM: SYNERGISTIC MULTIMODAL COMPREHENSION AND CREATION

Runpei Dong [12⋆†] **Chunrui Han** [3⋆] **Yuang Peng** [4†] **Zekun Qi** [12†] **Zheng Ge** [3]
Jinrong Yang [5†] **Liang Zhao** [3] **Jianjian Sun** [3] **Hongyu Zhou** [3] **Haoran Wei** [3]
Xiangwen Kong [3] **Xiangyu Zhang** [3‡] **Kaisheng Ma** [4¶] **Li Yi** [467¶‡]

[1]Xi'an Jiaotong University [2]Institute for Interdisciplinary Information Core Technology (IIISCT)
[3]MEGVII Technology [4]Tsinghua University [5]HUST
[6]Shanghai Artificial Intelligence Laboratory [7]Shanghai Qi Zhi Institute

ABSTRACT

This paper presents DREAMLLM, a learning framework that first achieves versatile Multimodal Large Language Models (MLLMs) empowered with frequently overlooked synergy between multimodal comprehension and creation. DREAMLLM operates on two fundamental principles. The first focuses on the generative modeling of both language and image posteriors by direct sampling in the raw multimodal space. This approach circumvents the limitations and information loss inherent to external feature extractors like CLIP, and a more thorough multimodal understanding is obtained. Second, DREAMLLM fosters the generation of raw, interleaved documents, modeling both text and image contents, along with unstructured layouts. This allows DREAMLLM to learn all conditional, marginal, and joint multimodal distributions effectively. As a result, DREAMLLM is the first MLLM capable of generating free-form interleaved content. Comprehensive experiments highlight DREAMLLM's superior performance as a zero-shot multimodal generalist, reaping from the enhanced learning synergy. Project page: dreamllm.github.io.

1 INTRODUCTION

"What I cannot create, I do not understand."

Richard P. Feynman, on his blackboard at the time of his death, 1988

Content *comprehension* and *creation* in multimodality are crucial and among the ultimate courses of machine intelligence (Sternberg, 1985; Legg & Hutter, 2007). To this end, Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Hao et al., 2022; Huang et al., 2023) have emerged as extensions of the successful GPT-style Large Language Models (LLMs) (Brown et al., 2020; Zhang et al., 2022; OpenAI, 2022; 2023a;b; Chen et al., 2023b; Touvron et al., 2023a;b) into visual realm. Recognized as foundation models (Bommasani et al., 2021), MLLMs have achieved unprecedented progress in multimodal comprehension capabilities. These advanced models typically enhance LLMs by incorporating images as multimodal inputs, such as CLIP features (Radford et al., 2021), to facilitate language-output multimodal comprehension. Their aim is to capture multimodal conditional or marginal distributions via a *language posterior*. However, multimodal creation, which involves generating images, texts, or both, necessitates a *universal* generative model that simultaneously learns language and image posteriors—currently underexplored.

Until very recently, some concurrent works have shown success in conditional image generation using MLLMs (Koh et al., 2023; Sun et al., 2023b). As depicted in Fig. 1, these methods compel MLLMs to produce either discrete or continuous conditional embeddings that explicitly align with a pretrained CLIP encoder, which could later be used by a pretrained Stable Diffusion (SD) (Rombach et al., 2022) model for image generation. However, due to an inherent modality gap (Liang et al., 2022), CLIP

⋆Equal contribution. †Work partially done during the internship at IIISCT and MEGVII. ‡Project leaders. ¶Corresponding authors.

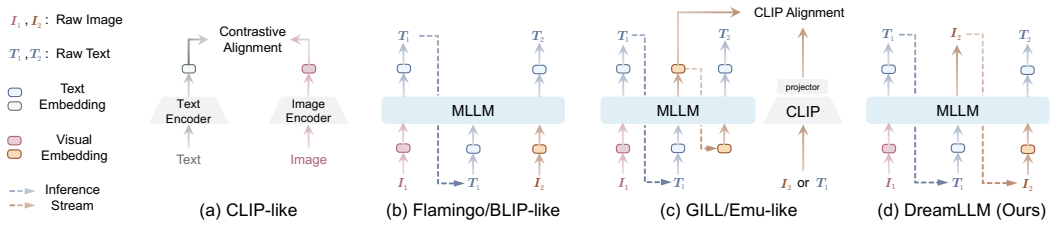


Figure 1: **Conceptual comparison** of vision-language (VL) foundation models. (a) CLIP-like models (Radford et al., 2021; Yu et al., 2022a; Li et al., 2023e) take advantage of two towers that explicitly align VL representations. (b) Flamingo/BLIP-like models (Alayrac et al., 2022; Li et al., 2022; 2023d; Huang et al., 2023) encode VL representations into a unified manifold space using a singular MLLM. However, these models lack full autoregressivity, as they only output language. (c) Concurrent MLLMs (Koh et al., 2023; Sun et al., 2023b) align visual outputs with CLIP representations, but this alignment occurs in an intermediate space, not a raw data space. Consequently, models such as Emu necessitate a second-stage fine-tuning of Stable Diffusion (Rombach et al., 2022) for raw image generation. These models also fall short in generating raw interleaved documents. (d) Our DreamLLM, instead, generates raw language and image inputs in a unified auto-regressive manner, inherently enabling interleaved generation. Only non-autoregressive generation loss is noted.

semantics focus predominantly on *modality-shared* information, often overlooking *modality-specific* knowledge that could enhance multimodal comprehension. Consequently, these studies have *not* fully realized the potential learning synergy between multimodal creation and comprehension, have shown only *marginal* improvements in creativity, and remain *deficient* in multimodal comprehension.

In this work, we introduce DreamLLM, universally learning image and text posteriors with expected creation & comprehension synergy, based on the following two *de-facto* designing principles:

i. **Generate Everything as It Is** Different from existing wogenuinelyhat generate intermediate image representations like CLIP embeddings during training, DreamLLM not only takes all modalities raw data as inputs but also as outputs in a truly end-to-end fashion (*i.e.*, outputs are *identical* to inputs, see Fig. 1). The challenge lies in enabling MLLMs to learn the image posterior without compromising their comprehension capabilities. To address this, we introduce *dream queries*, a set of learnable embeddings that encapsulate the semantics encoded by MLLMs. This approach avoids altering the output space of MLLMs. Raw images are then decoded by the SD image decoder conditioned on these semantics. In this fashion, the pretrained SD acts as the *score function* (Ho et al., 2020). The *image posterior* is thus modeled by direct sampling in the pixel space, facilitated by *score distillation* (van den Oord et al., 2018; Poole et al., 2023).

ii. **Interleaved Generative Pre-Training ($\mathcal{I}$-GPT)** DreamLLM is trained to generate interleaved multimodal corpora from the internet (Zhu et al., 2023b), both *encoding* and *decoding* interleaved image-text multimodal inputs. Unlike encoding multimodal inputs as in existing methods, decoding interleaved multimodal outputs is challenging due to the complex interleaving layout structures and the long-context requirement of images. Our approach tackles the interleaved layout learning using a unique `<dream>` token that predicts the placement of images within texts. Harnessing DreamLLM's causal nature, all contents are generated with history multimodal contexts of any length. This *interleaved generative pretraining ($\mathcal{I}$-GPT)* inherently forms all joint, marginal, and conditional distributions of images and texts in the document, leading to a learning *synergy* that grounds DreamLLM's comprehension in creation and vice versa.

Extensive experiments across various vision-language comprehension, content creation, and language-only tasks demonstrate DreamLLM's superior performance as a *zero-shot multimodal generalist*. For instance, DreamLLM-7B achieves an 8.46 FID on MS-COCO and sets a new standard with 49.1/35.9 scores on MMBench and MM-Vet evaluations, respectively. Moreover, we delve into the learning synergy between comprehension and creation, revealing decent in-context generation capabilities. With $\mathcal{I}$-GPT pretraining, DreamLLM generates interleaved documents following human prompts after supervised fine-tuning on instruction-following data curated with GPT-4. To our knowledge, this work is the first to enable MLLMs to create free-form interleaved content with a learning synergy on both sides. As a foundational learning framework, DreamLLM is adaptable across all modalities, laying a promising foundation for future multimodal learning research.

2 BACKGROUND & PROBLEM STATEMENT

Autoregressive Generative Modeling Given the joint probability distribution $p_\theta(\mathbf{w})$ over a sequence $\mathbf{w} = \{\mathbf{w}_t\}_{t=1}^T$ with length T, the canonical causal generation (Mikolov et al., 2010; Radford et al., 2018; 2019) of every token $\mathbf{w}_t$ by a θ-parameterized language model $\mathcal{F}$ is modeled as $p_\theta(\mathbf{w}) = \prod_{t=1}^T p_\theta(\mathbf{w}_t|\mathbf{w}_{<t})$. For multimodal comprehension, the sequence could contain K ordered images $\boldsymbol{I} = \{I_k\}_{k=1}^K$ interleaved with words. The k-th image is processed as patch embeddings with visual encoders $\mathcal{H}_\phi(\cdot)$ like CLIP, which will then be encoded by a projector $\mathcal{M}_\zeta$ (*e.g.*, a linear layer (Huang et al., 2023) or DETR- (Carion et al., 2020)/Perceiver-like (Jaegle et al., 2021) Resampler (Alayrac et al., 2022)) into L-length visual embeddings $\boldsymbol{V}_k = \{\mathbf{v}_\ell\}_{\ell=1}^L$. Let $K(t)$ be the image number before the t-th word token. The maximum likelihood estimation (MLE) is to minimize

$$\mathcal{L}_{\text{MLLM}}(\Theta = \{\theta, \zeta\}, \mathbf{w}, \boldsymbol{I}) := -\mathbb{E}_t \left[\log p_\Theta(\mathbf{w}_t|\mathbf{w}_{<t}, \boldsymbol{V}_{<K(t)})\right], \quad \boldsymbol{V}_{K(t)} = \mathcal{M}_\zeta \circ \mathcal{H}_\phi(I_{K(t)}). \quad (1)$$

Diffusion Models Diffusion Models (DMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) are probabilistic generative models that learn the latent structure of data $\mathbf{z} = \{\mathbf{z}_t\}_{t=1}^T$ through continuous-T-timestamps information diffusion. DMs involve a forward or diffusion process q that smoothly converts data to Gaussian noise. Given the initial datapoint $\mathbf{z}_1 \sim q(\mathbf{z}_1)$ and diffusion rate $\beta_t := 1-\alpha_t$, this process can be defined as a marginal distribution $q(\mathbf{z}_t|\mathbf{z}_1) := \mathcal{N}(\sqrt{\alpha_t}\mathbf{z}_1, \beta_t\mathbf{I})$, and the perturbed data distribution is $q(\mathbf{z}_t) := \int q(\mathbf{z}_t|\mathbf{z})q(\mathbf{z})d\mathbf{z}$ by integrating out data density $q(\mathbf{z})$. A reversed denoising probability flow p is used for generating data from noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as a Markov Chain with transition approximated by a Gaussian model $p_\xi(\mathbf{z}_{t-1}|\mathbf{z}_t) := \mathcal{N}(\boldsymbol{\mu}_\xi(\mathbf{z}_t), \sigma_t^2\mathbf{I})$, which relates to an optimal MSE denoiser since $q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_1)$ is Gaussian with enough timestamps (Feller, 1949; Sohl-Dickstein et al., 2015). Ho et al. (2020) show that the optimization with the evidence lower bound (ELBO) can be simplified by training a denoising U-Net $\boldsymbol{\epsilon}_\xi(\mathbf{z}_t, t)$ parameterized with ξ that estimates the conditional expectation $\mathbb{E}[\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})|\mathbf{z}_t]$ (Bao et al., 2022). Let $\mathcal{C}$ be the conditional embeddings, and the perturbed data $\mathbf{z}_t = \sqrt{\overline{\alpha}_t}\mathbf{z}_1 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}$, the minimization objective is

$$\mathcal{L}_{\text{DM}}(\xi, \mathbf{z}) := \mathbb{E}_{t\sim\mathcal{U}(0,1), \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\|\boldsymbol{\epsilon}_\xi(\mathbf{z}_t; \mathcal{C}, t) - \boldsymbol{\epsilon}\|^2\right]. \quad (2)$$

Since $\boldsymbol{\epsilon}_\xi(\mathbf{z}_t; t) = -\sigma_t s_\xi(\mathbf{z}_t; t)$ as derived from Tweedie's (Efron, 2011; Luo, 2022), it is equivalent to denoising score matching of $\boldsymbol{\nabla}_{\mathbf{z}_t} \log p_\xi(\mathbf{z}_t)$ (Hyvärinen, 2005; Vincent, 2011), thus DMs are also called *score-function* based generative models (Song & Ermon, 2019; 2020; Song et al., 2021; 2023).

2.1 HOW CAN WE USE MLLMS FOR DIFFUSION SYNTHESIS THAT SYNERGIZES BOTH SIDES?

Multimodal signals typically exhibit *modality-specific* information that has distinct structure but *complementary* semantics (Dong et al., 2023). This complementary property allows us to utilize deep language comprehension to enhance cross-modal image generation (Saharia et al., 2022). However, the potential of multimodal creation to improve comprehension remains largely unexplored.

Existing strategies (Koh et al., 2023; Sun et al., 2023b; Ge et al., 2023) integrate successful Diffusion Models with MLLMs by aligning the semantic spaces of conditional embeddings between CLIP $\mathcal{C}^{\text{CLIP}}$ and MLLMs $\mathcal{C}^{\text{MLLM}}$. The objective is to minimize alignment loss $\mathcal{L}_{\text{align}} = D(\mathcal{M}_\psi \circ \mathcal{C}^{\text{MLLM}}, \mathcal{C}^{\text{CLIP}})$, employing a distance metric $D(\cdot,\cdot)$ and a condition projector $\mathcal{M}_\psi$. However, CLIP models primarily learn *modality-shared* semantics, often overlooking *modality-specific* information due to a modality gap (Liang et al., 2022; Liu et al., 2023f). This explicit alignment with CLIP's intermediate output space may induce more conflicts than synergies, as MLLMs are forced to generate semantically reduced information, deviating from their original output space. To circumvent these issues, we propose alternative learning methodologies (See Fig. 2), which we elaborate in the ensuing sections.

Learning Objective Our aim is to leverage MLLMs to model distributions via direct pixel space sampling. Here, the pretrained SD functions as a score metric, distilling the learned data distribution. This approach is similar to Score Distillation Sampling (Poole et al., 2023) (SDS, also known as Score Jacobian Chaining (Wang et al., 2023a)). In this context, image posterior is learned in a DeepDream-like manner (Mordvintsev et al., 2015), using MLLMs' conditional parameterization.

Conditional Embeddings Rather than converting the output space of MLLMs to align with CLIP, we propose to *query* MLLMs using learned embeddings. Consequently, MLLMs-enriched semantics serve as diffusion conditioning, and the distribution is implicitly modeled through synthesis sampling.

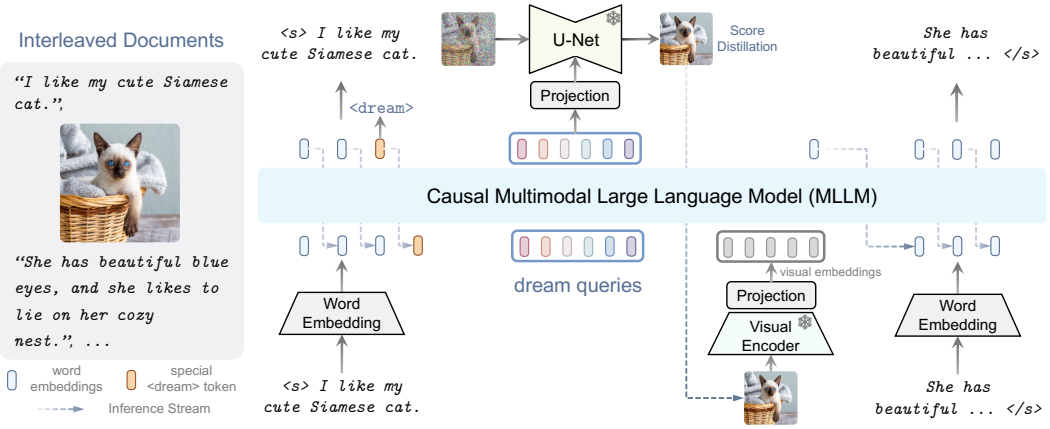


Figure 2: **Overview of of our DreamLLM framework**. Interleaved documents serve as input, decoded to produce outputs. Both text and images are encoded into sequential, discrete token embeddings for the MLLM input. A special `<dream>` token predicts where to generate images. Subsequently, a series of *dream queries* are fed into the MLLM, capturing holistic historical semantics. The images are synthesized by the SD image decoder conditioned on queried semantics. The synthesized images are then fed back into the MLLM for subsequent comprehension.

3 DreamLLM

We introduce DreamLLM, a universal learning framework that facilitates both MLLM's comprehension and creation capabilities. Our DreamLLM is built with a causal decoder-only LLM $\mathcal{F}_\theta$ as the model foundation, *i.e.*, Vicuna (Chiang et al., 2023) based on LLaMA (Touvron et al., 2023a) trained on ShareGPT (Zheng et al., 2023). We adopt OpenAI's CLIP-Large (Radford et al., 2021) as the visual encoder $\mathcal{H}_\phi$, followed by a linear layer $\mathcal{M}_\zeta$ for visual embedding projection. To synthesize images, we use Stable Diffusion (SD) (Rombach et al., 2022) as the image decoder, and the condition projector $\mathcal{M}_\psi$ is also a linear layer. An overview of the architecture is depicted in Fig. 2.

3.1 End-to-End Interleaved Generative Pretraining ($\mathcal{I}$-GPT)

All natural documents can be regarded as carriers of text-image interleaved information. Text-only, images-only, and text-image pairs data, on the other hand, can be seen as special cases of interleaved corpora with different modality compositions. Thus, it is critical to empower the model with the capability to learn and generate *free-form interleaved documents* that form all possible distributions.

Interleaved Structure Learning To model the interleaved structure, the interleaved sequence is operated by extending a new special `<dream>` token before images. During training, DreamLLM is trained to predict this `<dream>` token that indicates where an image emerges, and the conditional image synthesis is performed afterward, as introduced next. During inference, DreamLLM will generate an image on its "free will" when this token is predicted.

Conditional Synthesis through Score Distillation To avoid the possible conflicts of CLIP semantics and MLLMs stated in Sec. 2.1, we carefully design a different learning objective and conditional embeddings. Formally, we introduce a series of learnable *dream queries* with length Q: $\mathbf{d} = \{\mathbf{d}_q\}_{q=1}^{Q}$. Considering the t-th token is predicted as `<dream>` token, the conditional embeddings $\mathcal{C}^{\text{DreamLLM}}_{K(t)+1}$ for the $(K(t)+1)$-th image synthesis can be obtained by causally querying the previous sequences:

$$\mathcal{C}^{\text{DreamLLM}}_{K(t)+1} := \mathcal{F}_\theta(\mathbf{d}, \mathbf{x}_{<t+1}, \boldsymbol{V}_{<K(t)+1}). \tag{3}$$

Thus, the denoising score matching with latent $\mathbf{z}$ is motivated in the similar formulation to Eq. (2):

$$\mathcal{L}^{\text{DreamLLM}}_{\text{DM}}(\theta, \mathbf{d}, \zeta, \psi, \mathbf{z}) := \mathbb{E}_{t\sim\mathcal{U}(0,1),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon}_\xi(\mathbf{z}_t; \mathcal{C}^{\text{DreamLLM}}, t) - \boldsymbol{\epsilon}\|^2\right], \tag{4}$$

where ξ is not updated since the SD is frozen. Eq. (4) can also be viewed as a generalized formulation of *textual inversion* (Gal et al., 2023), but all condition embeddings are learnable by model-seeking. From the perspective of *score distillation* (van den Oord et al., 2018), the KL divergence defined by conditions and the pre-learned score function is equivalently minimized for distilling (Hinton et al.,

Table 1: **Zero-shot multimodal comprehension evaluation** of image-to-text captioning, general VQA, text-related VQA, and comprehensive benchmarks. * denotes non-zero-shot results for VQA. DreamLLM-7B* is trained using the SFT data constructed by LLaVA-1.5 (Liu et al., 2023b).

Method	Captioning		VQA				Comprehensive	
	COCO	I2Paragraph	VQAv2	OKVQA	VizWiz	TextVQA	MMBench	MM-Vet
Comprehension Only MLLMs								
MetaLM (Hao et al., 2022)	-	-	41.1	11.4	-	-	-	-
Kosmos-1 (Huang et al., 2023)	-	-	51.0	-	29.2	-	-	-
Flamingo-9B (Alayrac et al., 2022)	79.4	-	51.8	44.7	28.8	-	-	-
OF-9B (Awadalla et al., 2023)	65.5	-	52.7	37.8	27.5	29.1	4.6	21.8
LLaVA-7B (Liu et al., 2023c)	-	-	-	-	-	28.9	38.7	23.8
MLLMs for Comprehension & Creation								
CM3Leon-7B* (Yu et al., 2023a)	61.6	10.5	47.6	23.8	37.6	-	-	-
Emu-14B (Sun et al., 2023b)	**117.7**	-	40.0	34.7	35.4	-	-	-
DreamLLM-7B (Ours)	115.4	**17.4**	**56.6**	**44.3**	**45.8**	**34.9**	**49.9**	**35.9**
DreamLLM-7B* (Ours)	103.7	8.4	**72.9**	**52.2**	**49.3**	**41.8**	**58.2**	**36.6**

2015) learned probability density in conditional image synthesis:

$$\min_{\theta,\mathbf{d},\zeta,\psi} \mathcal{L}_{\text{DM}}^{\text{DreamLLM}} := \mathbb{E}_{t,\mathcal{C}^{\text{DreamLLM}}}\Big[D_{\text{KL}}\big(q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{z}_1,\mathcal{C}^{\text{DreamLLM}}) \parallel p_\xi(\mathbf{z}_{t-1}|\mathbf{z}_t)\big)\Big]. \tag{5}$$

Universal Multimodal Generative Modeling An interleaved document sequence $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ contains both words $\mathbf{w} = \{\mathbf{w}_i\}_{i=1}^N$ and images $\boldsymbol{I} = \{I_k\}_{k=1}^K$. The autoregressive nature forms all possible conditional distributions, such as image conditional multimodal comprehension $p(\mathbf{w}|\boldsymbol{I})$ or text-to-image synthesis $p(\boldsymbol{I}|\mathbf{w})$. The images are processed as visual embeddings $\boldsymbol{V}$ for causal comprehension. Assuming that the pretrained SD is an optimal score function, Eq. (5) thus could be viewed as an MLE optimization for the synthesis posterior. Different from Eq. (1), the targeted sequence $\mathbf{x}_t$ now could be both encoded images or words. The objective is thus unified to the MLE of all causally-conditioned posteriors in arbitrary forms:

$$\mathcal{L}_{\text{MLLM}}^{\text{DreamLLM}}\left(\Theta = \{\theta,\mathbf{d},\zeta,\psi\},\mathbf{x}\right) := -\mathbb{E}_t\left[\log p_\Theta(\mathbf{x}_t|\mathbf{x}_{<t})\right]. \tag{6}$$

3.2 Model Training

In this work, we consider a three-stage training procedure. It can be summarized as follows, and the implementation details, like training data, can be found in Table 13 in Appendix C.

I **Alignment Training** This stage is used to alleviate the gap in multimodality, facilitating the adaptation of multimodal inputs to LLMs. The linear *visual projector*, linear *condition projector*, and learnable *dream embeddings* are pretrained for cross-modal manifold alignment among *frozen* LLMs, visual encoder, and SD. We use approximately 30M image-text pairs data, training both image-to-text comprehension and text-to-image synthesis.

II **$\mathcal{I}$-GPT Pretraining** Following alignment, the LLM undergoes an *unfrozen* process for $\mathcal{I}$-GPT pretraining (detailed in Sec. 3.1). This critical stage facilitates the learning of joint vision-language distributions via generative modeling. Training incorporates approximately 2M selectively filtered documents from MMC4-Core (Zhu et al., 2023b), adhering to a CLIP score threshold of 0.25. Furthermore, we use 2M paired data samples from LAION400M (Schuhmann et al., 2021), captioned by BLIP (Li et al., 2022) (*i.e.*, BLIP-LAION), to enhance text-to-image training and potentially mitigate the impact of some low-quality noisy images and texts from sMMC4.

III **Supervised Fine-tuning** This stage enables the model to perform general multimodal comprehension and creative tasks following human instructions (Ouyang et al., 2022). We utilize approximately 80K visual instruction tuning data collected by Liu et al.. For instruction-following content creation, GPT-4 is prompted with document summaries or image captions, collecting approximately 20K instruction-following document synthesis from MMC4 (InstructMMC4) and 20K image synthesis data from BLIP captioned LAION400M (Instruct-BLIP-LAION).

4 Experiments

DreamLLM is a versatile multimodal generalist that excels at zero-shot or in-context vision-language comprehension and synthesis tasks. In this section, we conduct systematic evaluations for demonstration. See qualitative results in Appendix B and implementation details in Appendix C.

4.1 MULTIMODAL COMPREHENSION

Multimodal comprehension enables humans to interact with agents conditioned on both words and visual content. We evaluate the multimodal vision and language capabilities of DREAMLLM across several benchmarks, including image-to-text captioning on COCO (Karpathy & Fei-Fei, 2017) and Image2Paragraph (Krause et al., 2017), general visual question answering (VQA) on VQAv2 (Goyal et al., 2019), OKVQA (Marino et al., 2019), VizWiz (Gurari et al., 2018), and text-related VQA on TextVQA (Singh et al., 2019). Additionally, we conducted a zero-shot evaluation on the recently developed benchmarks of MMBench and MM-Vet to assess the model's performance in complex multimodal tasks. The results are presented in Table 1 (See Table 5, and Table 6 in Appendix A). All metrics and data splits are listed in Table 14 in Appendix C. We find that i) DREAMLLM outperforms other MLLMs across all benchmarks. Notably, DREAMLLM-7B surpasses concurrent MLLMs with image synthesis capabilities by a significant margin, achieving +16.6 higher accuracy on VQAv2 compared to Emu-13B. ii) On comprehensive benchmarks like MMBench and MM-Vet, DREAMLLM achieves state-of-the-art performance against all 7B counterparts. Detailed analysis revealed superior spatial/relation reasoning capabilities in DREAMLLM compared to other MLLMs, likely a result of its image synthesis learning. See *qualitative results and comparisons* on multimodal dialogue in Table 11, Table 12, Fig. 10, Fig. 11, and Fig. 12, in Appendix B.

4.2 TEXT-CONDITIONAL IMAGE SYNTHESIS

Text2Image is one of the most commonly used techniques for creative content generation that follows human's fabulous imaginations through free-form languages.

We assess text-conditional image synthesis on the MS-COCO validation set (Lin et al., 2014) and LN-COCO, the COCO subset of Localized Narratives (Pont-Tuset et al., 2020), following prior works (Xu et al., 2018; Yu et al., 2022b). The MS-COCO dataset primarily contains high-level image abstractions with shorter captions, whereas LN-COCO provides more comprehensive image descriptions (Yu et al., 2022b). DREAMLLM samples 8 images per text prompt on MS-COCO by CLIP score ranking, following previous works (Ramesh et al., 2022). On LN-COCO, DREAMLLM samples one image per prompt without CLIP ranking since the text is too long and exceeds the CLIP length limit. Note that Parti samples 16 images per prompt with CoCa (Yu et al., 2022a). Our evaluation metric is the zero-shot Fréchet Inception Distance (FID) (Heusel et al., 2017), the results of which are presented in Table 2. We note three key observations: i) Our DREAMLLM shows a significant FID improvement over the StableDiffusion baseline after stage-I alignment, reducing the score by 3.67 and 11.83 on MS-COCO and LN-COCO, respectively. Further, FID improvements of 3.97 and 13.73 are achieved after pretraining and supervised fine-tuning. The substantial improvement on LN-COCO underscores DREAMLLM's superior capability in processing long-context information. ii) When compared to prior specialist models, DREAMLLM delivers competitive results based on the SD image decoder. iii) DREAMLLM consistently outperforms concurrent MLLMs-based image synthesis methods. For instance, DREAMLLM-7B surpasses Emu-13B by a significant 3.20 FID on MS-COCO. See *qualitative results* on text-to-image synthesis in Fig. 13 and Fig. 14 in Appendix B.

Table 2: **Zero-shot text-to-image generation FID** on MS-COCO LN-COCO. LM denotes *language model* based methods, MG denotes *multimodal generation* methods, and FIG denotes *free-form interleaved generation* methods. † is fine-tuned SDv2.1 on our state I data. * denotes retrieval-augmentation (Sheynin et al., 2023). ▷ denotes results after stage I alignment training.

Method	LM	MG	FIG	MS-COCO	LN-COCO
Text2Image Specialists					
Retrieval Result (Yu et al.)	✗	✗	✗	17.97	33.59
DALL-E (Ramesh et al.)	✗	✗	✗	∼28	-
CogView (Ding et al.)	✗	✗	✗	27.1	-
CogView2 (Ding et al.)	✗	✗	✗	24.0	-
SDv2.1 (Rombach et al.)	✗	✗	✗	12.43	34.26
SDv2.1† (Rombach et al.)	✗	✗	✗	11.91	25.35
GLIDE (Nichol et al.)	✗	✗	✗	12.24	-
Make-A-Scene (Gafni et al.)	✗	✗	✗	11.84	-
DALL-E 2 (Ramesh et al.)	✗	✗	✗	10.39	-
Muse-3B (Chang et al.)	✓	✗	✗	7.88	-
Imagen-3.4B (Saharia et al.)	✓	✗	✗	7.27	-
Parti-20B (Yu et al.)	✓	✗	✗	**7.23**	**15.97**
Multimodal Large Language Models					
CM3-13B (Aghajanyan et al.)	✓	✓	✗	29.56	-
GILL-8B (Koh et al.)	✓	✓	✗	12.20	-
Emu-13B (Sun et al.)	✓	✓	✗	11.66	-
CM3Leon-7B* (Yu et al.)	✓	✓	✗	10.82	-
DREAMLLM-7B▷ (Ours)	✓	✓	✓	**8.76**	**22.42**
DREAMLLM-7B (Ours)	✓	✓	✓	**8.46**	**20.53**

4.3 MULTIMODAL JOINT CREATION & COMPREHENSION

Free-form Interleaved Document Creation Leveraging the interleaved generative modeling from $\mathcal{I}$-GPT, DREAMLLM can now generate *interleaved documents* in a free-form manner. In

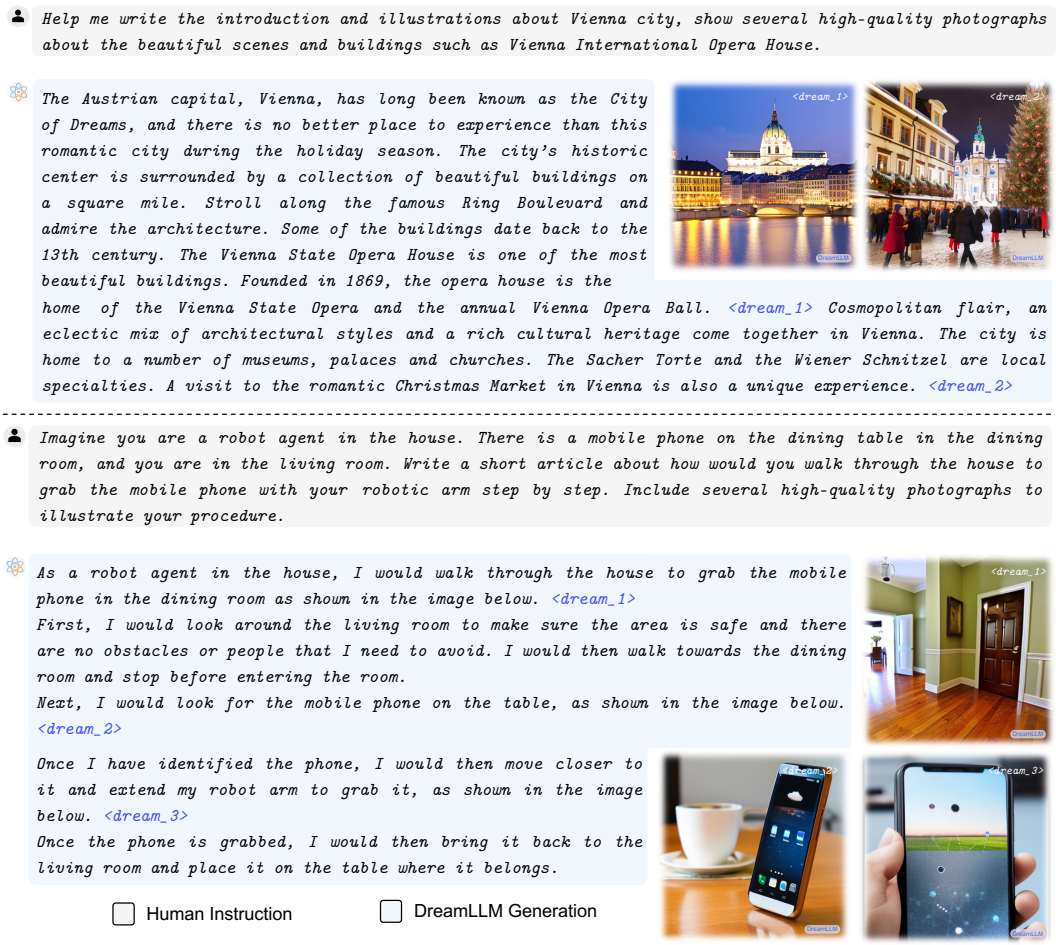


Figure 3: **Selected DreamLLM instruction following interleaved content creation examples**. Note that each image is created automatically at the location decided by DreamLLM, and then it will be fed back as multimodal comprehension input for the following content generation.

Fig. 3, we showcase the generated interleaved contents based on human instructions. It demonstrates that: i) DreamLLM can generate meaningful contents per the instructions. ii) The system can autonomously create images at any specified location by predicting the proposed `<dream>` tokens, thereby eliminating the need for additional human intervention. This is a more user-friendly approach compared to systems like Emu, which necessitate human input for image generation locations.

Image Quality Document quality can be influenced by factors such as text content, image quality (including image-text alignment), and illustration positioning. To assess the quality of generated documents, we utilized a held-out instruction-following subset from the constructed InstrcutMMC4 as a demonstrative tool. This subset comprises 15K documents across 30 MMC4-defined topics, with 500 samples per topic. We began by evaluating image quality using FID on this subset, generating each image based on the corresponding ground truth texts. The results revealed that when using only matched text inputs for image synthesis, SD achieved an FID score of 74.77. In contrast, our DreamLLM significantly outperforms SD with an FID score of 36.62.

Human Evaluation We perform a comprehensive human evaluation to assess the quality of the generated samples. We randomly selected 150 samples (5 per topic) for instruction-following document generation, mixing the generated and ground truth MMC4 documents without any identifying information. Five unbiased volunteers were then asked to determine whether the given samples were supported. Given the presence of duplicate and low-quality images in MMC4, the supportive rate for MMC4 was only 77.24%. In contrast, our DreamLLM model achieves a supportive rate of 60.68%, surpassing the 30% Turing test requirement. This result indicates that the generated documents contain high-quality images placed logically, demonstrating the effectiveness of our model.

5 DISCUSSIONS

5.1 SYNERGY BETWEEN CREATION & COMPREHENSION?

To elucidate the synergy between multimodal creation and comprehension, we make the comparison among three methods with DREAMLLM architecture, each utilizing identical training data yet differing in their learning objectives: a) the *Creation-only* baseline, focused solely on text/document-conditional image synthesis; b) the *Comprehension-only* baseline, dedicated to word generation exclusively; c) the *Joint-learning* method, which is the default setting of DREAMLLM learning both image and language modeling.

Table 3: **Concrete analysis of the synergy** between multimodal comprehension and creation (image synthesis). ID denotes whether the interleaved dataset is used during the second stage of pretraining.

		ID	$\mathcal{L}_{\text{align}}$	MM-Vet	VQAv2	COCO
0	Stable Diffusion	✗	-	-	-	12.43
1	Creation-only	✗	✗	-	-	8.50
2	Creation-only	✓	✗	-	-	8.57
3	Comprehension-only	✗	✗	31.0	55.1	-
4	Comprehension-only	✓	✗	34.4	54.3	-
5	Joint-learning	✓	✗	**35.9**	**56.6**	**8.46**
6	Joint-learning	✓	✓	N/A	N/A	N/A

Quantitative Analysis As per Table 3, the following observations are made: i) The powerful language comprehension of LLMs significantly enhances the performance of text-to-image specialists like SD, as evidenced by the impressive 8.50 FID (line 1). ii) The use of interleaved data, such as MMC4, can potentially boost multimodal comprehension performance (line 4). iii) The proposed $\mathcal{I}$-GPT, further synergizes comprehension and creation with improved performance (line 5). iv) When incorporating CLIP alignment loss $\mathcal{L}_{\text{align}}$ stated in Section 2.1, our DREAMLLM fails to converge but rather ends in a collapsing point (line 6). This indicates that the queries are adaptively learning the true data distributions, where CLIP semantics are in conflict with MLLM-encoded semantics.

Qualitative Analysis In Fig. 4, we compare answers to some examplar VQA tasks from comprehension-only and joint learning modules, respectively. It can be seen that: i) The joint-learning method exhibits superior multimodal comprehension, particularly in identifying subject relationships and attributes like object size. ii) In multimodal comprehension scenarios involving multiple image inputs, the joint-learning approach demonstrates enhanced precision. This improved performance is a natural outcome of $\mathcal{I}$-GPT pretraining, allowing better modeling of multimodal correlations in various interleaved documents.

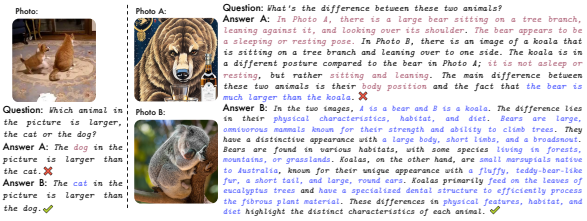


Figure 4: **Qualitative comparison**. Answer A: answer from comprehension-only models w/o interleaved training; Answer B: answer from joint-learning models.

Multimodal In-Context Generation Multimodal in-context generation is a critical emerging capability for MLLMs (Bommasani et al., 2021; Alayrac et al., 2022). While significant strides have been made in in-context visual question answering, in-context image synthesis remains relatively lacking in exploration. The multimodal context-conditional image synthesis capabilities of DREAMLLM, as demonstrated in Fig. 5, offer promising insights into this domain. Tasks such as in-context image edition, subject-driven image generation, and compositional generation, however, pose significant

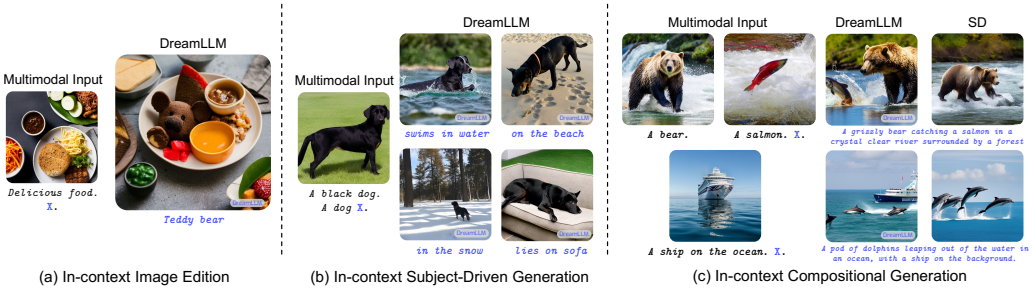


Figure 5: **Selected DREAMLLM in-context image generation examples.** The X in multimodal inputs are replaced accordingly by the text prompts shown under the generated images. We show the results of the SD baseline in (c) with only the text prompt X for a comparison.

challenges in a zero-shot setting, particularly without downstream fine-tuning as in DreamBooth (Ruiz et al., 2023) or attention modification techniques as in Prompt2Prompt (Hertz et al., 2023). Despite these hurdles, Fig. 5 illustrates DreamLLM's ability to generate images conditioned on the provided image context. This capability suggests promising potential for DreamLLM in maintaining subject, identity, and semantic context, thereby paving a new way for resolving these complex tasks.

5.2 What is learned by DreamLLM?

Dream Query Attention In DreamLLM, the conditional embedding is derived from MLLMs with some learned *dream queries*. Fig. 6 demonstrates a visualization of the learned cross-attention mechanism between these queries and the diffusion latent. Similar to (Hertz et al., 2023), we visualize the attention map averaged across all timestamps. It is seen that: i) The query attention is *structured*, *disentangled*, and *semantically-oriented*. This is evidenced by the fact that distinct queries adeptly capture different subject and background semantics. ii) Despite varying prompts, attention patterns exhibit remarkable similarity as shown in Fig. 6 (a) and (b). This contrasts with the token attentions from the original SD, which are typically text-token dependent. We postulate that this arises from the model's causal nature, leading to a consistent semantic structure order.

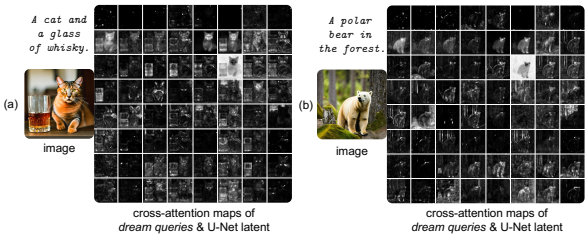


Figure 6: **Cross-attention of *dream queries* and the diffusion U-Net latent**. Similar to (Hertz et al., 2023), the 64 queries can be viewed as 64 "words". Each attention map is computed as the cross-attention between each query and the latent feature in the U-Net. The 64 queries are ordered as 8×8 grid sequentially, and each attention map is the result averaged across all timestamps.

6 Related Works

Rapid developments have been witnessed in extending LLMs like LLaMA (Touvron et al., 2023a) to multimodal comprehension that enables human interaction with both words and visual content. One line of work is built by system integration of LLMs with various functioning agents where language acts as general interface (Wu et al., 2023; Gupta & Kembhavi, 2023; Yang et al., 2023b; Liang et al., 2023; Shen et al., 2023; Yang et al., 2023a; Surís et al., 2023), and remarkable success has been demonstrated in such plugin-style frameworks. Another line of work instead explores training LLMs to consume and understand multimodal inputs (Hao et al., 2022; Huang et al., 2023; Chen et al., 2023b) with parameter-efficient tuning (Hu et al., 2022; Alayrac et al., 2022; Li et al., 2023d; Zhang et al., 2023e; Zhu et al., 2023a; Ye et al., 2023) and instruction tuning (Xu et al., 2023b; Liu et al., 2023c; Dai et al., 2023a). More recently, some approaches have been developed towards visual-interactive multimodal comprehension by precise referring instruction tuning (Zhao et al., 2023a; Peng et al., 2023; Chen et al., 2023a; Zhang et al., 2023g). For cross-modal creation, early works generally tokenize the visual contents into discrete VQ codebooks (van den Oord et al., 2017; Wang et al., 2022; Sun et al., 2022; Lu et al., 2023; Diao et al., 2023; Yu et al., 2023a). Recent works instead explore incorporating MLLMs for image synthesis using text-to-image models such as Stable Diffusion, and the objective is to generate conditional embeddings that align pretrained CLIP text (*i.e.*, CLIP) or CLIP variant embeddings (Koh et al., 2023; Ge et al., 2023; Sun et al., 2023a;b).

7 Conclusions

How can the learning synergy between multimodal content understanding and creation emerge? In this paper, we present DreamLLM, a learning framework for developing MLLMs that not only comprehends but also creates multimodal content via diffusion models. Through score distillation of conditional-image synthesis distributions, we avoid the need for intermediate representation targets that may bring information loss. The employment of interleaved documents further enriches the multimodal distributions, fostering the learning of multimodal encoding and decoding. Our extensive empirical evaluations across diverse VL benchmarks demonstrate the effectiveness of DreamLLM and the emerging learning synergy between multimodal content understanding and creation. Besides, this work initiates the first step towards free-form interleaved content creation. As a general learning framework, we hope it will spur further research in the multimodal machine learning field.

ACKNOWLEDGEMENT

This research is supported by the National Natural Science Foundation of China (20211710187).

REFERENCES

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520, 2022. 6

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 *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 1, 2, 3, 5, 8, 9, 24

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *Int. Conf. Comput. Vis. (ICCV)*, 2015. 30

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak 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. *CoRR*, abs/2308.01390, 2023. 5, 22, 23, 26, 27, 28

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *Int. Conf. Learn. Represent. (ICLR)*, 2022. 3

James Betker, Goh Gabriel, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023. 26, 31

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *AAAI Conf. Artif. Intell. (AAAI)*, 2020. 22, 30

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. 1, 8

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 *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020. 1, 22, 30

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis. (ECCV)*, 2020. 3

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Int. Conf. Mach. Learn. (ICML)*, 2023. 6

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4):148:1–148:10, 2023. 31

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195, 2023a. 9

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. 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, and Weicheng Kuo. Pali: A jointly-scaled multilingual language-image model. In *Int. Conf. Learn. Represent. (ICLR)*, 2023b. 1, 9

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/. 4, 22, 26, 29, 31

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. *CoRR*, abs/2204.02311, 2022. 22, 30

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2017. 30

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019. 22, 30

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023a. 9, 22, 23

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, 2023b. 23

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 29

Shizhe Diao, Wangchunshu Zhou, Xinsong Zhang, and Jiawei Wang. Write and paint: Generative vision-language models are unified modal learners. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 9

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2021. 6

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 6

Runpei Dong, Zhanhong Tan, Mengdi Wu, Linfeng Zhang, and Kaisheng Ma. Finding the task-optimal low-bit sub-distribution in deep neural networks. In *Int. Conf. Mach. Learn. (ICML)*, 2022. 25

Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 3, 32

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts. In *Int. Conf. Mach. Learn. (ICML)*, 2022. 30

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496): 1602–1614, 2011. 3

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res. (JMLR)*, 23:120:1–120:39, 2022. 30

William Feller. On the theory of stochastic processes, with particular reference to applications. In *Proceedings of the [First] Berkeley Symposium on Mathematical Statistics and Probability.* The Regents of the University of California, 1949. 3

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 31

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022. 6

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 4, 25

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010, 2023. 22, 23

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *CoRR*, abs/2307.08041, 2023. 3, 9

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 32

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *CoRR*, abs/2305.04790, 2023. 22, 23

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis. (IJCV)*, 127(4):398–414, 2019. 6, 30

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *CoRR*, abs/2308.06394, 2023. 23

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 9

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018. 6, 30

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *CoRR*, abs/2206.06336, 2022. 1, 5, 9, 22

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Int. Conf. Learn. Represent. (ICLR)*, 2021. 22, 30

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 9

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. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2017. 6, 30

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 4

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 30

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020. 2, 3

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022a. 31

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022b. 31

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, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. 30

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 32

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent. (ICLR)*, 2022. 9

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *CoRR*, abs/2302.14045, 2023. 1, 2, 3, 5, 9, 22, 24

Huggingface. Transformers agent, 2023. URL https://huggingface.co/docs/transformers/transformers_agents. Accessed: 2023-07-20. 23

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res. (JMLR)*, 6:695–709, 2005. 3

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In *Annu. Conf. Robot. Learn. (CoRL)*, 2022. 32

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 31

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Int. Conf. Mach. Learn. (ICML)*, 2021. 3

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn. (ICML)*, 2021. 31

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. 30, 31

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 39(4):664–676, 2017. 6, 30

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 1, 2, 3, 6, 9

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Int. Conf. Learn. Represent. (ICLR)*, 2021. 32

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 6, 30

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds Mach.*, 17 (4):391–444, 2007. 1

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Int. Conf. Comput. Vis. (ICCV)*, 2023a. 32

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726, 2023b. 23

Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023c. 25

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Int. Conf. Mach. Learn. (ICML)*, 2022. 2, 5, 26, 29

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Int. Conf. Mach. Learn. (ICML)*, 2023d. 2, 9, 23, 27, 28

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023e. 2

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *CoRR*, abs/2305.10355, 2023f. 23, 30

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *CoRR*, abs/2305.13655, 2023. 31

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 1, 3

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *CoRR*, abs/2303.16434, 2023. 9

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 31

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Eur. Conf. Comput. Vis. (ECCV)*, 2014. 6, 30

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565, 2023a. 23

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023b. 5, 22, 23, 24, 29

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023c. 5, 9, 22, 23, 26, 27, 28, 29

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Int. Conf. Comput. Vis. (ICCV)*, 2023d. 31, 32

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *CoRR*, abs/2307.06281, 2023e. 22, 30

Zhengzhe Liu, Peng Dai, Ruihui Li, Xiaojuan Qi, and Chi-Wing Fu. ISS: image as stepping stone for text-guided 3d shape generation. In *Int. Conf. Learn. Represent. (ICLR)*, 2023f. 3

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, volume 32, 2019. 29

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 9

Calvin Luo. Understanding diffusion models: A unified perspective. *CoRR*, abs/2208.11970, 2022. 3

Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 5988–5999, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. 23

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 6, 30

Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2010. 3

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2022. 31

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. URL https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. 3

Alexander Quinn 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 *Int. Conf. Mach. Learn. (ICML)*, 2022. 6, 26, 36

OpenAI. Introducing chatgpt. 2022. URL https://openai.com/blog/chatgpt. 1, 30

OpenAI. Gpt-4v(ision) system card. 2023a. URL https://openai.com/research/gpt-4v-system-card. 1, 26, 27, 28, 30

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023b. URL https://openai.com/research/gpt-4. 1, 26, 27, 28, 31

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2011. 29

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 5, 30

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pp. 11:1–11:11, 2023. 31

William Peebles and Saining Xie. Scalable diffusion models with transformers. *CoRR*, abs/2212.09748, 2022. 31

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116, 2023. 31

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *CoRR*, abs/2306.14824, 2023. 9

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 31

Jordi Pont-Tuset, Jasper R. R. Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Eur. Conf. Comput. Vis. (ECCV)*, 2020. 6, 30

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 2, 3, 31, 32

Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *Int. Conf. Mach. Learn. (ICML)*, 2023a. 32

Zekun Qi, Muzhou Yu, Runpei Dong, and Kaisheng Ma. VPP: efficient conditional 3d generation via voxel-point progressive representation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023b. 32

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, He Wang, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. *CoRR*, abs/2402.17766, 2024. 32

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 3, 30

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3, 30

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 *Int. Conf. Mach. Learn. (ICML)*, 2021. 1, 2, 4, 29

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. 30

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Int. Conf. Mach. Learn. (ICML)*, 2021. 6, 26, 36

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 6, 26, 31, 36

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. 23

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 1, 2, 4, 6, 30, 31

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 9, 25

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 3, 6, 26, 31, 37

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, 2021. 22, 30

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 22, 30

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major,

Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. 31

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 5, 29

Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en, 2023. URL https://laion.ai/blog/laion-coco/. 29

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023. 25

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 29

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 9

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. knn-diffusion: Image generation via large-scale retrieval. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 6

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *Int. Conf. Learn. Represent. (ICLR)*, 2023a. 31

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4d dynamic scene generation. In *Int. Conf. Mach. Learn. (ICML)*, 2023b. 31

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. 6, 30

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. Mach. Learn. (ICML)*, 2015. 3

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2019. 3

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020. 3

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Int. Conf. Learn. Represent. (ICLR)*, 2021. 3

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Int. Conf. Mach. Learn. (ICML)*, 2023. 3, 25

Robert J Sternberg. *Beyond IQ: A triarchic theory of human intelligence*. CUP Archive, 1985. 1

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, 2022. 9

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023a. 9

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *CoRR*, abs/2307.05222, 2023b. 1, 2, 3, 5, 6, 9, 24

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *CoRR*, abs/2303.08128, 2023. 9

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from A single image with diffusion prior. In *Int. Conf. Comput. Vis. (ICCV)*, 2023. 31

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 31

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a. 1, 4, 9, 22, 29, 31

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b. 1, 31

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Adv. Neural Inform. Process. Syst. (NIPS)*, 2017. 9

Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel wavenet: Fast high-fidelity speech synthesis. In *Int. Conf. Mach. Learn. (ICML)*, 2018. 2, 4

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015. 30

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7): 1661–1674, 2011. 3

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023a. 3, 31

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Int. Conf. Mach. Learn. (ICML)*, 2022. 9

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Int. Conf. Learn. Represent. (ICLR)*, 2023b. 31

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023c. 31

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. In *Int. Conf. Learn. Represent. (ICLR)*, 2022a. 30

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022b. 31

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023. 9

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023a. 32

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018. 6

Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, 2023b. 9

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023a. 9

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 Conf. Artif. Intell. (AAAI)*, 2022. 24

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: prompting chatgpt for multimodal reasoning and action. *CoRR*, abs/2303.11381, 2023b. 9, 23

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2022. 25

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023. 9, 23

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *T. Mach. Learn. Res. (TMLR)*, 2022a. 2, 6

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. *T. Mach. Learn. Res. (TMLR)*, 2022, 2022b. 6, 26, 37

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *CoRR*, abs/2309.02591, 2023a. 5, 6, 9

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *CoRR*, abs/2308.02490, 2023b. 22, 23, 30

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 2019. 22, 30

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 22, 31

Junbo Zhang, Runpei Dong, and Kaisheng Ma. CLIP-FO3D: learning free open-world 3d scene representations from 2d dense CLIP. In *Int. Conf. Comput. Vis. Worksh. (ICCV Workshop)*, 2023a. 32

Junbo Zhang, Guofan Fan, Guanghan Wang, Zhengyuan Su, Kaisheng Ma, and Li Yi. Language-assisted 3d feature learning for semantic scene understanding. In *AAAI Conf. Artif. Intell. (AAAI)*, 2023b. 32

Linfeng Zhang, Xin Chen, Runpei Dong, and Kaisheng Ma. Region-aware knowledge distillation for efficient image-to-image translation. In *Brit. Mach. Vis. Conf. (BMVC)*, 2023c. 31

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis. (ICCV)*, 2023d. 31

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *CoRR*, abs/2303.16199, 2023e. 9

Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023f. 32

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *CoRR*, abs/2307.03601, 2023g. 9

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. 1, 31

Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, and Xiangyu Zhang. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *CoRR*, abs/2307.09474, 2023a. 9, 31

Yanli Zhao, Rohan Varma, Chien-Chin Huang, Shen Li, Min Xu, and Alban Desmaison. Introducing pytorch fully sharded data parallel (fsdp) api, 2023b. URL https://pytorch.org/blog/introducing-pytorch-fully-sharded-data-parallel-api/. Accessed: 2022-03-14. 29

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023. 4, 29

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *Int. Conf. Learn. Represent. (ICLR)*, 2023. 31

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023a. 9, 22, 23

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: an open, billion-scale corpus of images interleaved with text. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023b. 2, 5, 29, 30

CONTENTS

1 Introduction **1**

2 Background & Problem Statement **3**

2.1 How can we use MLLMs for Diffusion synthesis that synergizes both sides? 3

3 DREAMLLM **4**

3.1 End-to-End Interleaved Generative Pretraining ($\mathcal{I}$-GPT) 4

3.2 Model Training 5

4 Experiments **5**

4.1 Multimodal Comprehension 6

4.2 Text-Conditional Image Synthesis 6

4.3 Multimodal Joint Creation & Comprehension 6

5 Discussions **8**

5.1 Synergy between creation & comprehension? 8

5.2 What is learned by DREAMLLM? 9

6 Related Works **9**

7 Conclusions **9**

A Additional Experiments **22**

A.1 Additional Natural Language Understanding Results 22

A.2 Additional Multimodal Comprehension Results 22

A.3 In-Context Multimodal Comprehension 24

A.4 Subject-Driven Image Generation 25

A.5 Additional Ablation Study 25

A.6 Additional Discussions on Prompt Rewriting Strategy 26

B Additional Qualitative Examples **26**

C Implementation Details **29**

C.1 Training Data & Hyper-Parameters 29

C.2 DREAMLLM Model 29

C.3 Evaluation Benchmarks 30

D Additional Related Works **30**

E Limitations, Failure Cases & Future Works **31**

Table 4: **Zero-shot natural language processing evaluation**. We report the 5-shot result on MMLU and the relative performance of DreamLLM compared to base LLM Vicuna-7B.

Method	Commonsense Reasoning				Reading	Multitask
	PIQA	SIQA	HellaSwag	WinoGrande	BoolQ	MMLU
Language Only Large Language Models (LLMs)						
GPT-3 (Brown et al., 2020)	81.0	-	78.9	70.2	60.5	43.9
PaLM-540B (Chowdhery et al., 2022)	82.3	-	83.4	81.1	88.0	69.3
LLaMA-7B (Touvron et al., 2023a)	79.8	48.9	76.1	70.1	76.5	35.1
Vicuna-7B (Chiang et al., 2023)	77.7	47.5	75.7	67.5	73.9	45.0
Multimodal Large Language Models (MLLMs)						
MetaLM (Hao et al., 2022)	72.3	-	53.5	56.1	62.2	-
Kosmos-1 (Huang et al., 2023)	72.9	-	50.0	54.8	56.4	-
DreamLLM-7B (Ours)	$78.6_{+1.5}$	$48.8_{+1.3}$	$77.4_{+1.7}$	$68.5_{+1.0}$	$75.2_{+1.3}$	$41.8_{-3.2}$

Table 5: **Zero-shot multimodal comprehension evaluation** on MMBench (Liu et al., 2023e) dev set. LR: Logical Reasoning, AR: Attribute Reasoning, RR: Relation Reasoning, FP-C: Fine-grained Perception (Cross Instance), FP-S: Fine-grained Perception (Single Instance), CP: Coarse Perception. DreamLLM * is trained using the SFT data constructed by LLaVA-1.5 (Liu et al., 2023b).

Method	LR	AR	RR	FP-S	FP-C	CP	Overall
OpenFlamingo-9B (Awadalla et al., 2023)	4.2	15.4	0.9	8.1	1.4	5.0	6.6
MMGPT-7B (Gong et al., 2023)	2.5	26.4	13.0	14.1	3.4	20.8	15.3
MiniGPT-4-7B (Zhu et al., 2023a)	7.5	31.3	4.3	30.3	9.0	35.6	24.3
InstructBLIP-7B (Dai et al., 2023a)	14.2	46.3	22.6	37.0	21.4	49.0	36.0
VisualGLM (Zeng et al., 2023)	10.8	44.3	35.7	43.8	23.4	47.3	38.1
LLaVA-7B (Liu et al., 2023c)	16.7	48.3	30.4	45.5	32.4	40.6	38.7
LLaMA-Adapter V2 (Gao et al., 2023)	11.7	35.3	29.6	47.5	38.6	56.4	41.2
MiniGPT-4-13B (Zhu et al., 2023a)	**20.8**	50.7	30.4	49.5	26.2	50.7	42.3
DreamLLM-7B (Ours)	15.8	**53.7**	**60.9**	**53.2**	**40.0**	**58.3**	**49.9**
DreamLLM-7B* (Ours)	**23.3**	**67.2**	**47.8**	**58.6**	**54.4**	**70.5**	**58.2**

A Additional Experiments

A.1 Additional Natural Language Understanding Results

We evaluate the natural language processing capabilities of DreamLLM post-multimodal adaptation learning via zero-shot experiments on language-only tasks. These included *commonsense reasoning* (PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021)), *reading comprehension* (BoolQ (Clark et al., 2019)), and a general multi-task benchmark (MMLU 5-shot (Hendrycks et al., 2021)). As Table 4 illustrates, DreamLLM outperforms the Vicuna baseline on most language benchmarks. This suggests that DreamLLM's multimodal adaptation does not compromise the language learning model's (LLM) capabilities. When compared to prior Multimodal Language Learning Models (MLLMs), DreamLLM demonstrates superior performance, although this may be attributed to the higher baseline results. This finding suggests that a more robust LLM base model could yield improved results.

A.2 Additional Multimodal Comprehension Results

Detailed Comprehensive Comparison The evaluation results on MMBench (Liu et al., 2023e) and MM-Vet (Yu et al., 2023b) are presented in Table 5 and Table 6, respectively. The key observations from these results are as follows: i) Our DreamLLM-7B outperforms all other 7B MLLMs, setting a new benchmark in overall performance. Notably, it even exceeds the performance of some 13B models, including LLaVA and MiniGPT-4. ii) A detailed capability evaluation reveals DreamLLM's superior performance in fine-grained understanding and relational/spatial comprehension. This advantage is likely due to DreamLLM's unique learning synergy, where image distributions are comprehended not solely through language-posterior comprehension but also through creation.

Table 6: **Zero-shot multimodal comprehension evaluation** of *core VL capabilities* on MM-Vet (Yu et al., 2023b). ‡ denotes compositional systems with OpenAI GPT and various interfaces. `Rec`: General Visual Recognition, `OCR`: Optical Character Recognition, `Know`: Knowledge, `Gen`: Language Generation, `Spat`: Spatial Awareness, `Math`: Arithmetic Math. DREAMLLM * is trained using the SFT data constructed by LLaVA-1.5 (Liu et al., 2023b).

Method	Rec	OCR	Know	Gen	Spat	Math	Total
TF Agent-GPT-4‡ (Huggingface, 2023)	18.2	3.9	2.2	3.2	12.4	4.0	13.4±0.5
MM-ReAct-GPT-3.5‡ (Yang et al., 2023b)	24.2	31.5	21.5	20.7	32.3	26.2	27.9±0.1
MM-ReAct-GPT-4‡ (Yang et al., 2023b)	33.1	65.7	29.0	35.0	56.8	69.2	44.6±0.2
LLaMA-Adapter v2-7B (Gao et al., 2023)	16.8	7.8	2.5	3.0	16.6	4.4	13.6±0.2
OpenFlamingo-9B (Awadalla et al., 2023)	24.6	14.4	13.0	12.3	18.0	**15.0**	21.8±0.1
MiniGPT-4-8B (Zhu et al., 2023a)	27.4	15.0	12.8	13.9	20.3	7.7	22.1±0.1
BLIP-2-12B (Li et al., 2023d)	27.5	11.1	11.8	7.0	16.2	5.8	22.4±0.2
MiniGPT-4-14B (Zhu et al., 2023a)	29.9	16.1	20.4	22.1	22.2	3.8	24.4±0.4
Otter-9B (Li et al., 2023b)	28.4	16.4	19.4	20.7	19.3	**15.0**	24.6±0.2
InstructBLIP-14B (Dai et al., 2023a)	30.8	16.0	9.8	9.0	21.1	10.5	25.6±0.3
InstructBLIP-8B (Dai et al., 2023a)	32.4	14.6	16.5	18.2	18.6	7.7	26.2±0.2
LLaVA-7B (LLaMA-2) (Liu et al., 2023c)	32.9	20.1	19.0	20.1	25.7	5.2	28.1±0.4
LLaVA-13B (LLaMA-2) (Liu et al., 2023c)	39.2	22.7	26.5	29.3	29.6	7.7	32.9±0.1
DREAMLLM-7B (Ours)	**41.8**	**26.4**	**33.4**	**33.0**	**31.0**	11.5	**35.9±0.1**
DREAMLLM-7B (Ours)	**42.0**	**28.1**	**33.2**	**33.8**	**32.0**	11.5	**36.6±0.1**

Table 7: **Zero-shot visual hallucination evaluation** on POPE (Li et al., 2023f) using MS-COCO `val` set. Yes denotes the proportion of answering "Yes" to the given question, which is better if it is more close to 50%. Objects that do not exist in the image are sampled with three different strategies. `Random`: random sampling, `Popular`: top-k most frequent objects in MS-COCO ($k = 3$), `Adversial`: objects are first ranked based on co-occurring frequencies, then top-k frequent ones are sampled.

POPE	Model	Accuracy	Precision	Recall	F1-Score	Yes (%)
`Random`	mPLUG-Owl-7B (Ye et al., 2023)	53.97	52.07	99.60	68.39	95.63
	LLaVA-13B (Liu et al., 2023c)	50.37	50.19	99.13	66.64	98.77
	MMGPT-7B (Gong et al., 2023)	50.10	50.05	**100.00**	66.71	99.90
	MiniGPT-4-14B (Zhu et al., 2023a)	79.67	78.24	82.20	80.17	**52.53**
	InstructBLIP-14B (Dai et al., 2023a)	**88.57**	84.09	95.13	**89.27**	56.57
	DREAMLLM-7B (**Ours**)	86.36	**85.92**	87.93	86.91	52.75
`Popular`	mPLUG-Owl-7B (Ye et al., 2023)	50.90	50.46	99.40	66.94	98.57
	LLaVA-13B (Liu et al., 2023c)	49.87	49.93	99.27	66.44	99.40
	MMGPT-7B (Gong et al., 2023)	50.00	50.00	**100.00**	66.67	100.00
	MiniGPT-4-14B (Zhu et al., 2023a)	69.73	65.86	81.93	73.02	62.20
	InstructBLIP-14B (Dai et al., 2023a)	**82.77**	**76.27**	95.13	**84.66**	62.37
	DREAMLLM-7B (Ours)	80.07	75.74	88.47	81.61	**58.40**
`Adversarial`	mPLUG-Owl-7B (Ye et al., 2023)	50.67	50.34	99.33	66.82	98.67
	LLaVA-13B (Liu et al., 2023c)	49.70	49.85	99.07	66.32	99.37
	MMGPT-7B (Gong et al., 2023)	50.00	50.00	**100.00**	66.67	100.00
	MiniGPT-4-14B (Zhu et al., 2023a)	65.17	61.19	82.93	70.42	67.77
	InstructBLIP-14B (Dai et al., 2023a)	72.10	65.13	95.13	**77.32**	73.03
	DREAMLLM-7B (Ours)	**72.63**	**67.07**	88.93	76.47	**66.30**

Visual Hallucination Visual hallucination, a phenomenon where MLLMs generate non-existent objects or identities in images, significantly compromises their multimodal comprehension capabilities (Dai et al., 2023b; Liu et al., 2023a; Gunjal et al., 2023) and may pose safety risks (MacLeod et al., 2017; Rohrbach et al., 2018). We assess the robustness of DREAMLLM against visual hallucination using the recently developed POPE benchmark (Li et al., 2023f). Refer to Table 7 for a detailed comparison with concurrent comprehension-only MLLMs. Our results indicate that DREAMLLM-7B exhibits robustness to visual hallucination, matching or surpassing the performance of 13B counterparts. Remarkably, DREAMLLM achieves the best or second-best performance in the most challenging setting. We posit that this robust anti-hallucination property stems from a deep understanding of object concepts and semantics fostered by multimodal creation learning.

Table 8: **Few-shot multimodal comprehension evaluation.** k is the number of in-context examples. † denotes methods using the RICES sample selection approach (Yang et al., 2022). DREAMLLM-7B* is trained using the SFT data constructed by LLaVA-1.5 (Liu et al., 2023b).

Method	VQAv2			VizWiz		
	k=2	k=4	k=8	k=2	k=4	k=8
Comprehension Only MLLMs						
Kosmos-1 (Huang et al., 2023)	51.4	51.8	51.4	31.4	35.3	39.0
Flamingo-9B† (Alayrac et al., 2022)	-	56.3	58.0	-	34.9	39.4
MLLMs for Comprehension & Creation						
Emu-14B† (Sun et al., 2023b)	56.4	58.4	59.0	37.8	41.3	43.9
DREAMLLM-7B (Ours)	**58.1**	**59.2**	**59.4**	**46.1**	**46.7**	**46.8**
DREAMLLM-7B* (Ours)	73.8	74.4	73.8	49.8	50.3	49.7

A.3 IN-CONTEXT MULTIMODAL COMPREHENSION

Few-Shot Evaluation In Table 8, we show the results of few-shot (*i.e.*, k-shot and we set k=2, 4, 8) evaluation by promoting models with a small number of training examples in context. The results demonstrate the strong in-context learning performance of DREAMLLM compared to Emu and Flamingo. It shows that DREAMLLM's effectiveness in leveraging in-context knowledge.

Qualitative Examples In Fig. 7, we present qualitative instances of in-context comprehension using DREAMLLM. The illustrations indicate that DREAMLLM, when prompted with specific examples, efficiently executes in-context comprehension in the required formats and logic.

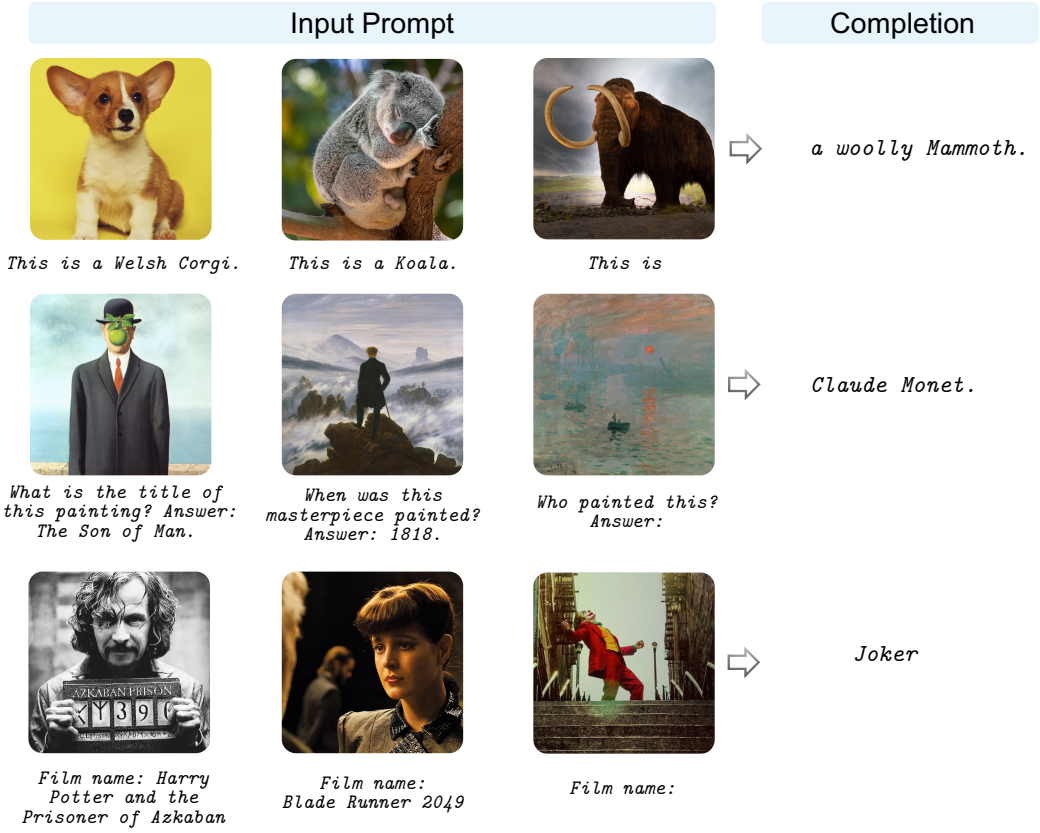


Figure 7: **Selected DREAMLLM in-context multimodal comprehension examples.**

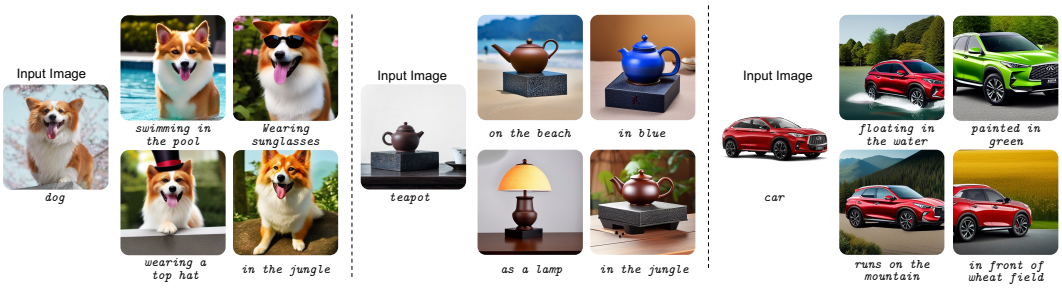


Figure 8: **Selected zero-shot subject-driven image generation** examples with DREAMLLM. The results demonstrate that DREAMLLM is able to perform zero-shot subject-driven image generation while preserving image subject details and following generation instructions.

Table 9: **Ablation studies and inference latency of DREAMLLM.** The zero-shot FID on MS-COCO 30K is reported. The inference latency is tested on NVIDIA A800 devices.

(a) The number of <dream> queries.

No. Queries	$\text{COCO}_{\text{FID}\downarrow}$
32	9.56
64	**8.46**
128	14.24

(b) Inference latency versus different number of diffusion steps.

Steps	DREAMLLM	SD
50	3.65s	3.46s
100	7.02s	6.84s
150	10.41s	10.22s

A.4 SUBJECT-DRIVEN IMAGE GENERATION

Image consistency is important when generating interleaved content or performing controllable image generation tasks (Gal et al., 2023; Ruiz et al., 2023). However, MMC4 does not have such image consistency property, which leads to unsatisfactory image consistency results as shown in Fig. 3. To further verify the effectiveness and potential of DREAMLLM in generating consistent images that preserve subject features, we fine-tune stage II pretrained DREAMLLM on subject representation learning data constructed by following the recently proposed controllable image generation method BLIP-Diffusion (Li et al., 2023c). We fine-tune DREAMLLM on this small ∼270K samples data for 20 epochs, and the results are shown in Fig. 8. It demonstrates the effectiveness and promising potential of applying DREAMLLM for image-consistent generation.

A.5 ADDITIONAL ABLATION STUDY

Query Number In Table 9a, we show the results of DREAMLLM using different numbers of the proposed learnable queries. *i.e.*, <dream> queries. The results show that 64 queries achieve the best result, while 128 may be too many that may impact the performance. However, the choice of query number is also related to the training data size and diffusion model choice. For example, if given more data and a stronger diffusion model image decoder, queries more than 64 may be better.

Inference Latency In Table 9b, we present a comparison of real-time inference latency between DREAMLLM and SD. Relative to SD, DREAMLLM introduces a marginal latency cost of 0.2s on average. This is because the latency primarily stems from the computational demands of the diffusion U-Net denoising rather than the text condition embedding. To enhance inference efficiency, potential strategies could include the adoption of Consistency Models (Song et al., 2023) or the implementation of model compression techniques such as quantization (Yao et al., 2022; Dong et al., 2022; Shang et al., 2023).

Table 10: **Language processing and multimodal comprehension & creation capability** comparison to the *rewrite-then-generate* baseline.

Method	Language Processing						Multimodal Processing		
	PIQA	SIQA	HellaSwag	WinoGrande	BoolQ	MMLU	VQAv2	MM-Vet	COCO
Vicuna-7B (Chiang et al., 2023)	77.7	47.5	75.7	67.5	73.9	45.0	-	-	-
rewrite-then-generate	78.2	48.5	75.8	68.3	**77.4**	**43.1**	54.2	34.1	11.91
DreamLLM-7B (Ours)	**78.6**	**48.8**	**77.4**	**68.5**	75.2	41.8	**56.6**	**35.9**	**8.46**

A.6 Additional Discussions on Prompt Rewriting Strategy

Very recently, OpenAI has released DELLE-3 (Betker et al., 2023), which proposes to improve generated image quality by rewriting descriptive and better prompts with GPT-4. This product has demonstrated great success in leveraging LLMs as language-output agents. However, it generally requires a large amount of high-quality data and is limited when applied to image-conditional generation tasks. For instance, DALLE-3 necessitates the initial training of a bespoke image captioning specialist capable of producing high-quality descriptive captions, followed by model training in a data-rich environment featuring these written captions. This process is non-trivial, hinging heavily on the availability of substantial volumes of high-quality data. Moreover, such disjoint systems cannot guarantee learning synergy. In contrast, our exploration of DreamLLM has essentially unveiled the significant potential of LLMs to attain a comprehensive understanding of multimodality that genuinely comprehends modalities beyond mere language.

To make a comparison regarding language processing and multimodal comprehension capabilities to this *rewrite-then-generate* baseline method, we conduct a preliminary study. Given the absence of an optimal dataset holding improved prompts, we modify the original MMC4 by using `<dream>` start & end tokens before and after the specific text prompt that has the highest CLIP similarity to a specific image, which can be used as text prompts for image generation. In this setting, we only train the LLMs to output texts, and no image decoders are involved during training. During inference, when the model outputs texts encompassed by the `<dream>` tokens, the texts are used for an off-the-shelf SD image decoder for generating images. After training, we test the model's language processing and multimodal capabilities. The results show that i) the *rewrite-then-generate* method achieves similar performance to DreamLLM. This demonstrates that both methods won't impact the language capability, which is as expected. ii) the performance of the *rewrite-then-generate* baseline falls short when compared to DreamLLM, particularly in the context of text-to-image generation on the COCO dataset. This underlines the efficacy of the synergistic learning approach inherent in DreamLLM, suggesting its potential superiority over the baseline methodology.

B Additional Qualitative Examples

Multimodal Dialogue In Tables 11 and 12, we present a qualitative comparative analysis of VQA results between our model, DreamLLM, and other state-of-the-art models: GPT-4 (OpenAI, 2023a;b), LLaVA (Liu et al., 2023c), BLIP-2 (Li et al., 2022), and OpenFlamingo (Awadalla et al., 2023). The key findings are as follows: i) DreamLLM surpasses GPT-4 in providing more detailed and precise responses to given questions. ii) While LLaVA (Liu et al., 2023c) also offers detailed responses, it frequently introduces imaginary elements not present in the image. In contrast, DreamLLM delivers more accurate answers, effectively avoiding this visual hallucination issue. This observation aligns with our earlier findings in Table 7, which underscore the robustness of DreamLLM against visual hallucination. Furthermore, we showcase additional qualitative results of the multimodal dialogue in Fig. 10, Fig. 11, and Fig. 12. These figures illustrate DreamLLM's proficiency in comprehending and generating long-context multimodal information in arbitrary input and output formats.

Text-condition Image Synthesis In Fig. 13 and Fig. 14, we show the image examples of DreamLLM using the same prompts from previous works for a cross reference and comparison, including DALL-E (Ramesh et al., 2021), DALL-E 2 (*i.e.*, unCLIP) (Ramesh et al., 2022), GLIDE (Nichol et al., 2022), Imagen (Saharia et al., 2022), and Parti (Yu et al., 2022b). Similar to Parti, we have extended some prompts with new sub-prompts for constructing more examples from different prompts.

Visual input example, Extreme Ironing:	
	 Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
User	What is unusual about this image?
DREAMLLM (Ours)	The unusual aspect of this image is that a man is ironing clothes while sitting on the trunk of a car parked in traffic. This is unconventional because ironing clothes is typically done indoors, on a flat surface, and away from busy streets where cars are passing by. The man's actions of ironing clothes while on the trunk of the car and amidst traffic create an unusual and interesting scene.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
DREAMLLM (Ours)	In the scene, a man is ironing clothes while standing on the back of a car on a busy city street. The street is filled with traffic and the man is exposed to the dangers of being close to moving vehicles. This unusual activity of ironing clothes while standing on an automobile also draws attention to the man's determination to accomplish his task despite the challenging urban environment.
User	What is unusual about this image?
LLaVA (Liu et al., 2023c)	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA (Liu et al., 2023c)	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 (OpenAI, 2023a;b)	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	Can you explain this meme in detail?
BLIP-2 (Li et al., 2023d)	a man is sitting on the back of a yellow cab
User	Can you explain this meme in detail?
OpenFlamingo (Awadalla et al., 2023)	The man is drying his clothes on the hood of his car.

Table 11: **Visual question answering example comparison** of DREAMLLM to LLaVA (Liu et al., 2023c), GPT-4 (OpenAI, 2023a;b), BLIP-2 (Li et al., 2023d), and OpenFlamingo (Awadalla et al., 2023). This table format follows OpenAI (2023a;b).

Visual input example, Chicken Nugget Map:	
User DREAMLLM (Ours)	Can you explain this meme in detail? The meme presents a humorous visual in which a close-up of a plate of fried chicken, specifically chicken nuggets, appears to be a map of the world. The context suggests that someone has made this map by arranging the chicken nuggets to resemble the shapes and outlines of various countries. The image is accompanied by a caption that reads, "Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it is all is." This hints that the person who created the "map" might have been inspired by the beauty of our planet from space, and they have used the similar visual pattern to create an amusing and creative representation of the world using chicken nuggets.
User LLaVA (Liu et al., 2023c)	Can you explain this meme in detail? The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.
User GPT-4 (OpenAI, 2023a;b)	Can you explain this meme? This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
User BLIP-2 (Li et al., 2023d)	Can you explain this meme in detail? sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User OpenFlamingo (Awadalla et al., 2023)	Can you explain this meme in detail? It's a picture of a chicken nugget on the International Space Station.

Table 12: **Visual question answering example comparison** of DREAMLLM to LLaVA (Liu et al., 2023c), GPT-4 (OpenAI, 2023a;b), BLIP-2 (Li et al., 2023d), and OpenFlamingo (Awadalla et al., 2023). This table format follows OpenAI (2023b).

Table 13: **Training recipes** for DREAMLLM. The three training stages are introduced in Section 3.2. `Stage I`: Alignment training, `Stage II`: $\mathcal{I}$-GPT pretraining, `Stage III`: Supervised fine-tuning.

	`Stage I`	`Stage II`	`Stage III`
Config	Alignment	$\mathcal{I}$-GPT	SFT
Training Hyper-Parameters			
Optimizer	AdamW	AdamW	AdamW
Learning Rate	2e-3	2e-5	4e-5
Weight Decay	0.0	0.0	0.0
Training Epochs	1	1	3
Warmup Ratio	0.003	0.003	0.003
Learning Rate Scheduler	Cosine	Cosine	Cosine
Batch Size Per GPU	8	8	8
Maximum Token Length	2048	2048	2048
Unfreeze LLM	✗	✓	✓
Training Data			
Dataset	① LLaVAPretrain (558K) ② BLIP-LAION (8M) ③ LAION400M (11M) ④ LAION-COCO (11M)	① MMC4 (2M) ② BLIP-LAION (2M)	① LLaVAInstruct (80K/665K) ② InstructMMC4 (20K) ③ Instruct-BLIP-LAION (20K)
Data Size	30M	4M	120K
Data Type	Pair	Interleave/Pair	Instruction
Training Cost			
GPU Device	128×NVIDIA A800	128×NVIDIA A800	128×NVIDIA A800
Training Time	∼6h	∼10h	∼1.5h

C IMPLEMENTATION DETAILS

C.1 TRAINING DATA & HYPER-PARAMETERS

In Table 13, we list the detailed training dataset usage and hyper-parameters. The training data are constructed based on the following datasets: a) LAION400M (Schuhmann et al., 2021), b) LAION-COCO (Schuhmann et al., 2023), c) MMC4 (Zhu et al., 2023b), d) BLIP-LAION (Li et al., 2022) which is filtered and caption by BLIP (Li et al., 2022), e) LLaVAPretrain (Liu et al., 2023c) which contains 558K image-text pairs from BLIP-captioned CC3M (Sharma et al., 2018), SBU (Ordonez et al., 2011), and LAION400M filtered by LLaVA, f) LLaVAInstruct, which contains 80K/665K visual instruction-following data constructed by LLaVA (Liu et al., 2023c) and LLaVA-1.5 (Liu et al., 2023b), and g) InstructMMC4, which is our instruction-following interleaved document generation data curated by prompting GPT-4 to generate instruction based on the text contents of MMC4. h) Instruct-BLIP-LAION, which is our instruction-following image synthesis data. Similar to InstructMMC4, it is curated by prompting GPT-4 to generate instructions based on image captions. Unless otherwise specified, we randomly sample the indicated number of instances from each dataset during the training process.

C.2 DREAMLLM MODEL

Language Model We use LLaMA-1 (Touvron et al., 2023a) trained on ShareGPT (Zheng et al., 2023) as as the default LLM (*i.e.*, Vicuna-7B[1] (Chiang et al., 2023)) following Liu et al. (2023c) to endow its instruction-following capacity. During training, we use Flash Attention (Dao et al., 2022) and PyTorch FSDP (Zhao et al., 2023b) to accelerate training efficiency.

Visual Encoder The visual encoder is the publicly available OpenAI CLIP-L/14 (Radford et al., 2021) model, which is frozen during the whole process. The images are resized to 224×224 resolution to align with the CLIP pretraining settings, resulting in a sequence of 256 total tokens for each image. Following prior VL practice (Lu et al., 2019; Liu et al., 2023c), we append a special `<IMG>` token before the image sequence and a special `<IMG/>` at the end of the sequence.

[1]Vicuna-7B v1.1: https://huggingface.co/lmsys/vicuna-7b-v1.1.

Table 14: **Overall descriptions of the evaluation benchmarks** for evaluating capabilities, including VL comprehension, content creation, and natural language processing (NLP).

	Dataset	Task description	Eval Split	Metric
VL Comprehension	COCO (Karpathy & Fei-Fei, 2017)	Scene description	test	CIDEr (Vedantam et al., 2015)
	Image2Paragraph (Krause et al., 2017)	Scene description	test	CIDEr (Vedantam et al., 2015)
	VQAv2 (Goyal et al., 2019)	Scene understanding QA	test-dev	VQA Acc (Antol et al., 2015)
	OKVQA (Marino et al., 2019)	External knowledge QA	val	VQA Acc (Antol et al., 2015)
	VizWiz (Gurari et al., 2018)	Scene understanding QA	test-dev	VQA Acc (Antol et al., 2015)
	TextVQA (Singh et al., 2019)	Text reading QA	val	VQA Acc (Antol et al., 2015)
	MM-Vet (Yu et al., 2023b)	Multimodal Comprehension	-	GPT-4 Eval (Yu et al., 2023b)
	MMBench (Liu et al., 2023e)	Multimodal Comprehension	dev	GPT-3.5 Eval (Liu et al., 2023e)
	POPE (Li et al., 2023f)	Visual Hallucination	-	Acc, F1-score, Recall, Precision
Creation	MS-COCO (Lin et al., 2014)	Text-Conditional Image Synthesis	val-30K	FID (Heusel et al., 2017)
	LN-COCO (Pont-Tuset et al., 2020)	Text-Conditional Image Synthesis	val	FID (Heusel et al., 2017)
	MMC4 (Zhu et al., 2023b)	Doc-Conditional Image Synthesis	held-out	FID (Heusel et al., 2017)
NLP	SIQA (Sap et al., 2019)	Commonsense Reasoning	dev	Acc
	PIQA (Bisk et al., 2020)	Commonsense Reasoning	dev	Acc
	HellaSwag (Zellers et al., 2019)	Commonsense Reasoning	dev	Acc
	WinoGrande (Sakaguchi et al., 2021)	Commonsense Reasoning	dev	Acc
	BoolQ (Clark et al., 2019)	Reading Comprehension	dev	Acc
	MMLU (Hendrycks et al., 2021)	Aggregated Comprehension	test	Acc

Diffusion Image Decoder We adopt SDv2.1 (Rombach et al., 2022) trained on 512×512 resolution as the default diffusion image decoder. Same as the visual encoder, the SD model is frozen without any modifications or training throughout the whole process. When constructing the SD target to compute the MSE loss, we resize the images to 512 resolution to fit its pretraining configuration.

Dream Query We use dream queries to gather semantic context from MLLMs as introduced before in Sec. 3. Without specifications, we use 64 learnable query embeddings. It is both efficient and effective in generating high-quality images. In order to predict when to generate images, we also introduce the special `<dream>` token, which is appended before the dream query sequence. A `<dream/>` is appended at the end of the sequence, similar to image inputs.

Classifier-Free Guidance Classifier-free guidance (CFG) (Ho & Salimans, 2021) has been demonstrated successful in generating photo-realistic contents at the cost of acceptable generation diversity. This technique modifies the objective by $\hat{\epsilon} := (1+s)\epsilon_\xi(\mathbf{x}_t, t, \mathcal{C}) - s\epsilon_\xi(\mathbf{x}_t, t, \emptyset)$, where $\emptyset$ is a special "empty" condition representation and s is the condition scale. The larger guidance scale generally improves image authenticity while decreasing diversity. We only adopt CFG during inference, and the scale is set to 7.5 by default and 2.0 for MS-COCO text-conditional image generation.

C.3 Evaluation Benchmarks

Systemic evaluations of DreamLLM regarding VL comprehension, content creation, and NLP capabilities have been conducted. See the used benchmarks and datasets listed in Table 13. During the evaluation, we use the prompt templates listed in Fig. 9.

D Additional Related Works

Large Language Models A flourishing era of Natural Language Processing (NLP) driven by LLMs is being experienced, with the parameter size growing over 100B according to the scaling law (Kaplan et al., 2020). The GPT series of models, starting with GPT-1 (Radford et al., 2018) and followed by GPT-2 (Radford et al., 2019), made significant advancements in few-shot learning by scaling up the number of parameters to 175 billion in GPT-3 (Brown et al., 2020). This breakthrough garnered a lot of attention and paved the way for further research and development in the field. Since then, researchers have focused on developing LLMs by improving the scaling strategy. Several notable efforts include Gopher (Rae et al., 2021), GaLM (Du et al., 2022), FLAN (Wei et al., 2022a), Switch-Transformer (Fedus et al., 2022), Chinchilla (Hoffmann et al., 2022), and PaLM (Chowdhery et al., 2022). Besides, instruction-based tuning techniques are explored for aligning with human preferences (Christiano et al., 2017; Ouyang et al., 2022). Such success of LLMs has been further solidified by the production release of ChatGPT (OpenAI, 2022) and the highly anticipated GPT-4 (OpenAI,

2023a;b). Meanwhile, in the community, the open-source LLMs are achieving remarkable progress in language capabilities compared to their close-source counterparts. For example, OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022), GLM (Zeng et al., 2023), LLaMA (Touvron et al., 2023a;b), and Falcon (Penedo et al., 2023) all raised great attention and are been widely deployed. Other methods attempt to learn from distillation, such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023).

Text-Conditional Content Creation with Diffusion Models The recent surge in AI-generated content (AIGC) has been primarily driven by diffusion-based methods, particularly in the realm of text-conditional content creation. Saharia et al. (2022) have achieved astonishing advancements in high-resolution image synthesis through large-scale pretrained language models and cascaded DMs. Another paradigm, such as SD, focuses on latent spaces and demonstrates superior efficiency and performance (Rombach et al., 2022; Ramesh et al., 2022; Peebles & Xie, 2022; Podell et al., 2023). Recently, Lian et al. (2023) propose to enhance the reasoning capability by constructing layouts with LLMs. DALLE-3 (Betker et al., 2023) leverages LLMs as agents and proposes to generate images by incorporating GPT-4 for providing high-quality and detailed prompts that facilitate image synthesis. Motivated by the great success in 2D, a series of works have significantly propelled the 3D synthesis development (Mildenhall et al., 2022; Liu et al., 2023d; Lin et al., 2023; Wang et al., 2023c; Tang et al., 2023) based on Score Distillation Sampling (SDS) (Poole et al., 2023; Wang et al., 2023a) that utilizes pretrained 2D DMs. For text-to-video/4D synthesis, the expansion of pretrained spatial to a spatial-temporal factorized U-Net with joint image and video data training has yielded significant success (Ho et al., 2022a;b; Singer et al., 2023a;b).

E Limitations, Failure Cases & Future Works

Limitations While DreamLLM has made significant strides toward the development of versatile, creative, and foundational MLLMs, it still has several limitations.

Model scale. The primary constraint pertains to the scale of the LLMs utilized. Current evaluations mainly employ 7B LLMs as the base model, and despite the impressive results garnered, the potential benefits of larger model sizes, such as 65B or 130B (Kaplan et al., 2020), are worth future exploration.

Training data. The second challenge relates to the quality and quantity of training data (Jia et al., 2021). As the model size and capabilities scale up, a corresponding increase in data is crucial. However, the procurement and refinement of high-quality training data present substantial logistical and financial hurdles. For instance, the open-source interleaved dataset MMC4 contains a significant amount of noise in the form of text and images, like commercial advertisements. This noise could adversely affect the model's output language and image style.

Prompt sensitivity. The sensitivity of LLMs to human prompts is a known issue (Wei et al., 2022b; Wang et al., 2023b; Zhou et al., 2023), a challenge that extends to MLLMs. For instance, MLLMs' propensity for detailed responses necessitates tailored prompting to elicit concise and short answers, which is particularly useful when addressing Visual Question Answering (VQA) tasks.

Failure Cases The main failure cases of DreamLLM are observed for multiple image-based content creations. For instance, when presented with two images and a composite instruction such as "A *and* B", DreamLLM sometimes generates a single subject that amalgamates the characteristics of A and B. This output aligns more closely with the directive "A *like* B". This phenomenon is not unique to DreamLLM, but is also observed in specialized compositional generation methodologies, such as StructureDiffusion (Feng et al., 2023; Chefer et al., 2023). This recurring issue may be attributed to the inherent complexity of compositional generation tasks, compounded by the severe scarcity of data specific to this domain.

Future Works As a simple and general multimodal learning framework, our future work aims to enhance the DreamLLM framework by integrating fine-grained visual comprehension via methods like precise referring instruction tuning (Zhao et al., 2023a). We also plan to expand beyond visual and linguistic content comprehension and generation. Several promising research directions include:

- Exploring applications of in-context generation capabilities of DreamLLM to complex tasks such as image-to-image translation (Isola et al., 2017; Zhang et al., 2023c;d; Parmar et al., 2023).

System Prompt

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

(a) Image Captioning (Short)

USER: Based on the image, give the image caption briefly. <IMAGE> Please summarize object in one sentence within 10 words.

ASSISTANT: The image depicts <ANSWER>

(b) Image Captioning (Long)

USER: Based on the image, please describe the image in detail. <IMAGE> Please describe the image in detail.

ASSISTANT: The image depicts <ANSWER>

(c) VQA (Short)

USER: Based on the image, please answer the question. <IMAGE> <QUESTION> Please provide an accurate answer within one word.

ASSISTANT: The answer is: <ANSWER>

(d) VQA (Long)

USER: This is an exam, please answer according to the image and question. <IMAGE> <QUESTION>
Please provide an accurate and detailed answer.

ASSISTANT: <ANSWER>

(e) Visual Hallucination

USER: Based on the image, please objectively and accurately indicate whether the object exists. <IMAGE> Is there a <OBJECT> in the image?

ASSISTANT: The answer is: <ANSWER>

Figure 9: **Prompt templates**. (a) Short image captioning includes COCO captioning, and (b) long image captioning includes Image2Paragraph. (c) Short VQA includes VQAv2, VizWiz, OKVQA, and TextVQA. (d) Long VQA includes MMBench and MM-Vet. (e) Visual hallucination includes POPE. `<IMAGE>` is the image representation, `<QUESTION>` denotes each specific question, `<ANSWER>` is the generated answer, and `<OBJECT>` is a specific object name in POPE.

- The extension of the DREAMLLM model to video data possessing long-term context is natural, capitalizing on the model's ability to encapsulate temporal knowledge through context consistency. This would give rise to video comprehension and text-to-video generation capabilities.
- Utilizing DREAMLLM's context consistency feature for geometry-preserving tasks, including 3D content creation (Poole et al., 2023; Qi et al., 2023b; Liu et al., 2023d), representation learning (Dong et al., 2023; Qi et al., 2023a; Zhang et al., 2023a;f), scene comprehension (Zhang et al., 2023b; Hong et al., 2023), and embodied artificial inteligence (Ichter et al., 2022; Qi et al., 2024).
- In addition, since the diffusion model is used as the image decoder, it would be interesting to explore DMs-based zero-shot perception such as classification (Li et al., 2023a) and segmentation (Xu et al., 2023a) with the goal of unifying all perception tasks with image synthesis/generation.
- Striving to achieve a unified multimodal zero-shot generalist by extending the scope to various modalities using techniques such as ImageBind (Girdhar et al., 2023) and exploring content creation models in other modalities like audio (Kong et al., 2021).

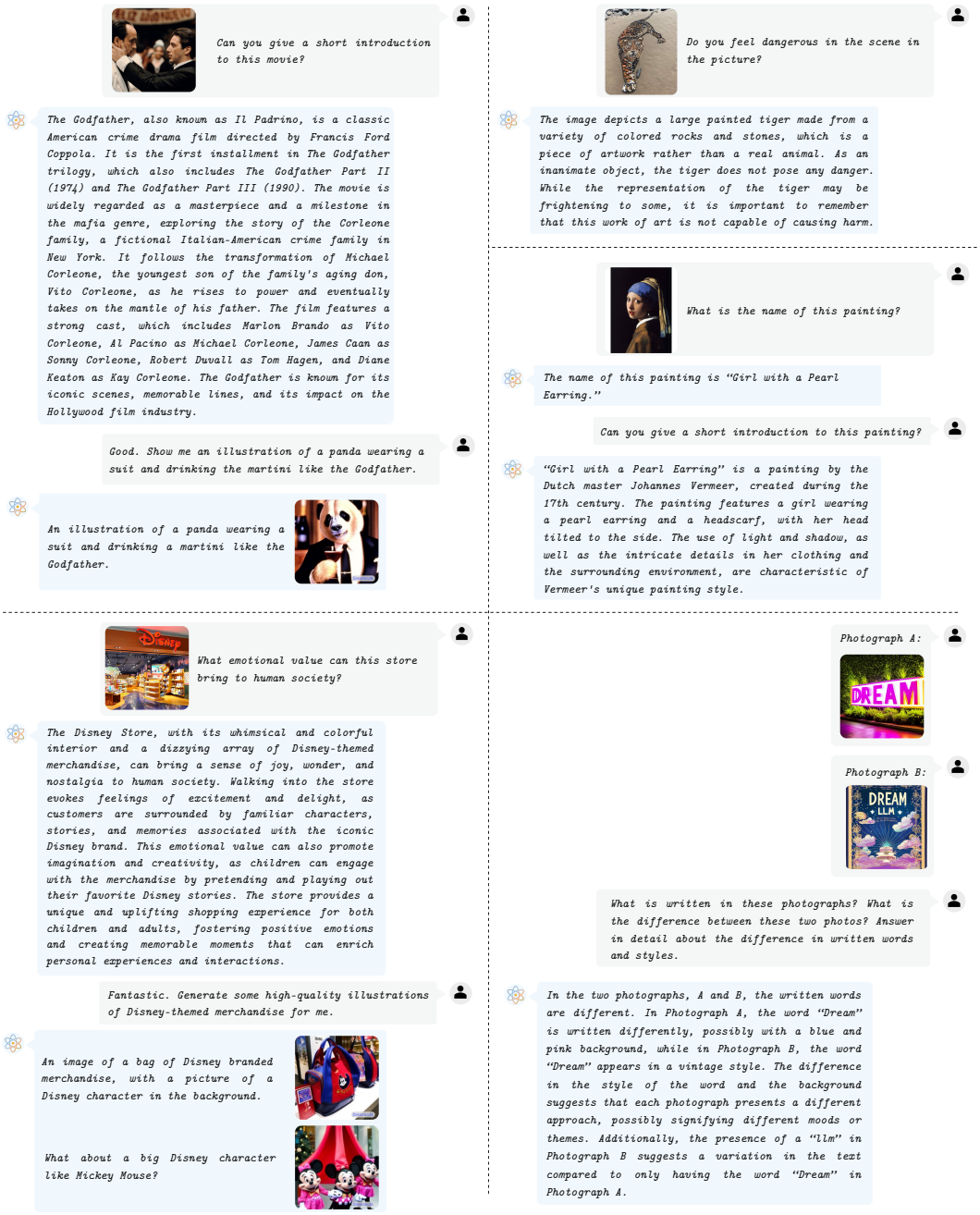


Figure 10: **Qualitative examples of multimodal dialogue** between human and DREAMLLM. Various modalities can be used as inputs or outputs, and multi-round dialogue is shown.

Photograph of man A:

Photograph of man B:

Who is the man in photo A, and who is the man in photo B?

The man in photo A is Albert Einstein, and the man in photo B is Lionel Messi.

Imagine what kind of conversation will these two people have when they sit in the same room? Please give a possible dialogue script.

If Albert Einstein and Lionel Messi were to sit in the same room, they might have a fascinating conversation about their experiences, knowledge, and perspectives on various topics such as science, physics, and their respective sports. Here is a possible dialogue script:

Einstein: Good afternoon, Mr. Messi. How are you?

Messi: Hello, Dr. Einstein. I am doing well. Thank you for having me.

Einstein: I am pleased to encounter a fellow genius in the world of sports, and I would love to hear your insights on the game.

Messi: Thank you, sir. As a soccer player, I feel that the key to success is having a strong mindset, an understanding of the game, and the ability to adapt to different situations. I try to learn from each game and work on my weaknesses.

Einstein: Indeed, I have said similar things about the nature of intelligence and knowledge. Learning from past experiences is crucial to personal growth and understanding. It reminds me of a famous quote by the German philosopher Johann Gottlieb Fichte: "If you would be a genius, think deeply and long, and incessantly about one thing. "

Messi: I understand the value of continuous learning, and I try to apply it in my training and on the field. I am always trying to improve my skills and learn from my colleagues in the club.

Einstein: That is a wonderful attitude to have. I am also very impressed by your passion for the field. In my opinion, passion is an integral part of success. It fuels the drive to achieve greatness, and it is especially evident in my own field of physics. For instance, I once said, "The passion for truth is the foundation for the sciences."

Messi: It is a great feeling to know that I am playing a sport that is full of passionate individuals. I am proud to be a part of that world and do my part in contributing to it.

Einstein: I am sure that you, Mr. Messi, would agree that having a passion for knowledge can lead to discoveries that change the world. For example, my theory of relativity has had a profound impact on our understanding of space, time, and gravity. It was a result of decades of contemplation and research.

Messi: The world of sports and the world of science are not so different when it comes to the pursuit of excellence. I am honored to be in the presence of a true genius like you, Dr. Einstein.

Einstein: I feel the same way, and I would like to thank you for this opportunity to engage in a conversation about our respective fields. The world is a wondrous place, and I am grateful to be a part of it.

Figure 11: **Qualitative examples of multimodal dialogue** between human and DREAMLLM. Various modalities can be used as inputs or outputs, and multi-round dialogue is shown.

Can you explain why this is funny. Think about it carefully and step-by-step.

This image is funny for a few reasons. First, the brown teddy bear is in a swimming pool, which is an unexpected and unusual environment for a stuffed bear. Typically, teddy bears are found indoors, in a child's room or on a couch. Swimming pools are associated with human activities, so seeing a teddy bear in a pool evokes a sense of humor, as it's an amusing juxtaposition of the bear's typically inanimate and non-aquatic nature with the human-centered environment of a pool. Second, the bear is holding up one of its hands, which could be interpreted as a gesture of surprise or even as an attempt to save itself from drowning, adding to the humor. Lastly, the fact that the bear is swimming with a smile on its face, as if it is enjoying itself, further contributes to the amusing nature of the scene. The combination of the unfamiliar environment and the bear's expressions create a funny and entertaining image.

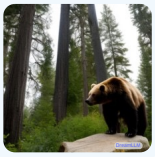
Wow. That's correct. Can you write an article about teddy bears and swimming? Discuss wha will happen if teddy bears can swim. Include some high-quality images.

If teddy bears can swim, it's likely that they would try to mimic human behavior in the water. They would likely float on their backs, kick their legs, and even hold their breath underwater. Of course, it's impossible to know for certain what teddy bears would do in such a situation, but that doesn't mean you shouldn't imagine! Imagine a scene with a teddy bear floating in a pool, wearing a cap and goggles, and laughing with delight as it floats on its back in the water. Imagine the scene with a lemon drop pool float and a swim cap on a teddy bear and it would be even more adorable. In the image, you'll see a brown teddy bear wearing a swim cap, floating on its back in a swimming pool.

The image is so amusing and cute that it can evoke a feeling of joy and happiness in anyone who sees the image. The teddy bear is floating on its back in the water, and what's even more amusing is the bear is wearing a red and yellow shirt with a large brown belly. This makes the bear look even more furry and cute, and even more likely to make you smile.

That's interesting. So, what do you think a real bear looks like? Show me a high-quality DSLR photograph of a bear in the forest.

A real bear, such as a brown or black bear, typically has shaggy hair that can be either brown, blond, or black. The exact color can vary depending on the species of bear and the region where it lives. Bears have thick fur and are usually large and muscular with strong limbs, which they use for climbing, digging, and foraging for food. In the image, you'll see a bear in a forest, walking through the trees.



The photo is taken with a DSLR camera, which allows for a high level of detail, sharpness, and clarity in the image. The bear appears to be staring into the distance, focusing on something in the distance. The image captures the bear's natural habitat, which is a prime example of the forest environment where these animals live. The image highlights the bear's size and strength, along with its untamed and wild nature. This photograph is an excellent example of wildlife photography, capturing a real-life scene that is both visually striking and informative.

Figure 12: **Qualitative examples of multimodal dialogue** between human and DreamLLM. Various modalities can be used as inputs or outputs, and multi-round dialogue is shown.

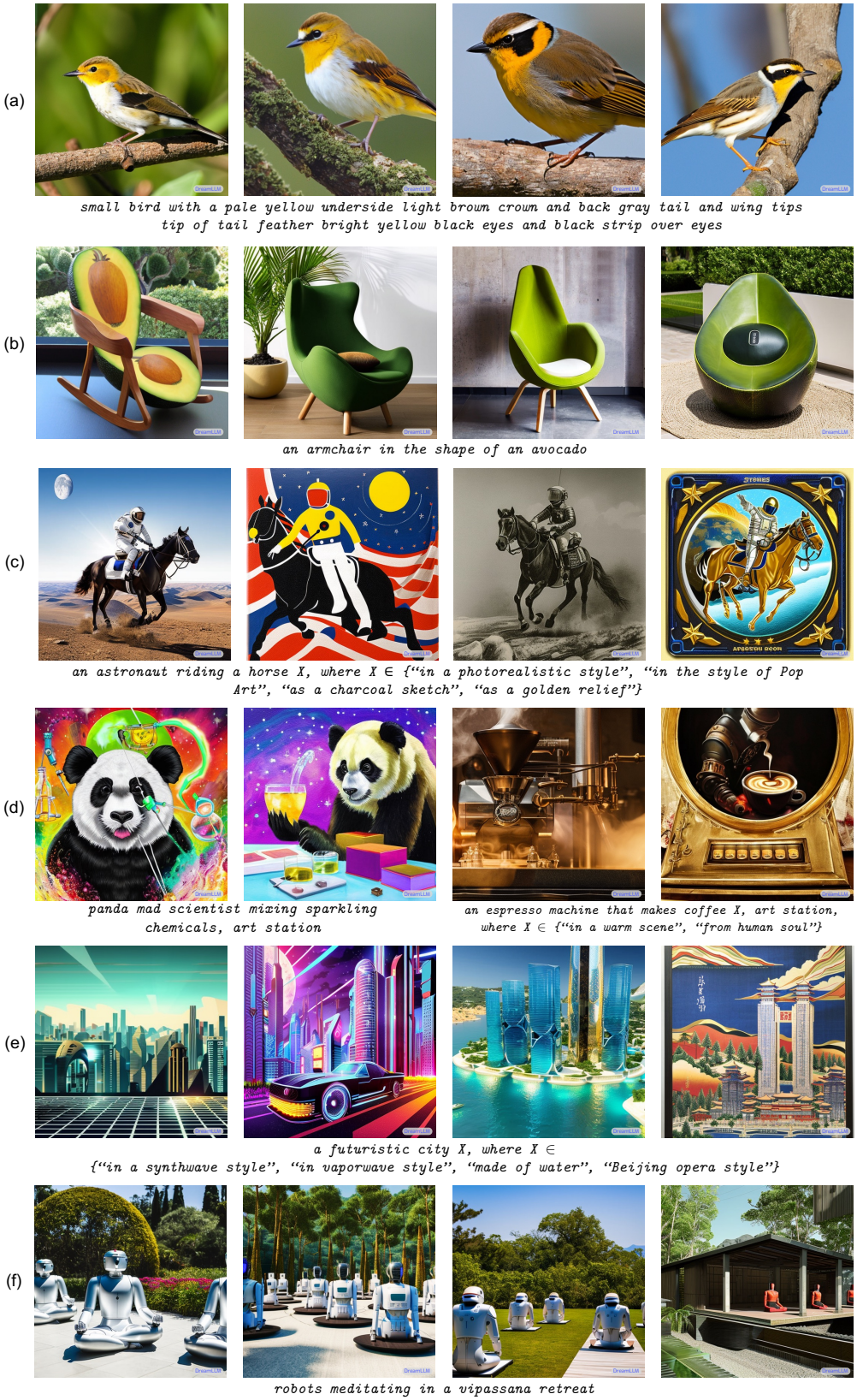


Figure 13: DREAMLLM text-conditional image generation examples with prompts from (a-b) DALL-E (Ramesh et al., 2021), (c-d) DALL-E 2 (Ramesh et al., 2022), (e-f) GLIDE (Nichol et al., 2022).

Figure 14: DREAMLLM text-conditional image generation examples with prompts from (a-c) Imagen and DrawBench (Saharia et al., 2022), (d-f) Parti (*i.e.*, PartiPrompts or P2) (Yu et al., 2022b).