This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MetaMorph: Multimodal Understanding and Generation via Instruction Tuning

Shengbang Tong[1,2] David Fan[1] Jiachen Zhu[1,2] Yunyang Xiong[3] Xinlei Chen[1] Koustuv Sinha[1]
Michael Rabbat[1] Yann LeCun[1,2] Saining Xie[2] Zhuang Liu[1,4]
[1]FAIR, Meta [2]New York University [3]Meta Reality Labs [4]Princeton University

Abstract

*In this work, we propose Visual-Predictive Instruction Tuning (**VPiT**)—a simple and effective extension to visual instruction tuning that enables a pretrained LLM to quickly morph into an unified autoregressive model capable of generating both text and visual tokens. VPiT teaches an LLM to predict discrete text tokens and continuous visual tokens from any input sequence of image and text data curated in an instruction-following format. Our empirical investigation reveals several intriguing properties of VPiT: (1) visual generation ability emerges as a natural byproduct of improved visual understanding, and can be unlocked efficiently with a small amount of generation data; (2) while we find understanding and generation to be mutually beneficial, understanding data contributes to both capabilities more effectively than generation data. Building upon these findings, we train our **MetaMorph** model and achieve competitive performance on both visual understanding and generation. In visual generation, MetaMorph can leverage the world knowledge and reasoning abilities gained from LLM pretraining, and overcome common failure modes exhibited by other generation models. Our results suggest that LLMs may have strong "prior" vision capabilities that can be efficiently adapted to both visual understanding and generation with a relatively simple instruction tuning process.*

1. Introduction

Multimodal Large Language Models (MLLMs) have advanced considerably in visual understanding, progressing from basic image captioning to complex visual inferences [4, 14, 36]. These models process multimodal inputs—primarily images and language—and generate text tokens. Multimodal LLMs often leverage a pretrained vision encoder [16, 52], a pretrained language model [3, 66], and align these modalities through connectors such as MLP [36, 37] or cross-attention modules [4, 14]. Among MLLM training methods, visual instruction tuning [36] has become widely used [2, 67]. It treats output embeddings of pretrained vision encoders as continuous-valued "visual tokens" and directly feeds them as inputs to pretrained LLMs.

One benefit of visual instruction tuning is that it is data and compute efficient. A pretrained LLM can be repurposed as a Multimodal LLM by instruction tuning with modest compute and data on the order of millions of image-text question-answer pairs [32, 63]. The effectiveness of visual instruction tuning indicates that LLMs already possess a considerable amount of inherent visual knowledge which allows them to learn and develop visual understanding during the instruction tuning process [85]. Inspired by this, we investigate whether LLMs can also be finetuned to *generate* visual information with comparable efficiency and effectiveness.

Current attempts toward "unified" models—models capable of both multimodal understanding and generation—often treat visual generation as an orthogonal capability to visual understanding. They tend to require substantial changes to the original MLLM architecture and significant multimodal pretraining and/or finetuning. Designing such methods is challenging, and past research takes different approaches including tokenizing visual inputs into discrete tokens [39, 62, 73], incorporating diffusion objectives [75, 86], and decoupling vision into separate understanding and generation modes [71]. For example, approaches like LWM [39], Show-o [75], and Chameleon [62] require billions of image-text pairs [17, 56] for extensive pretraining and finetuning.

In this work, we propose *Visual-Predictive Instruction Tuning* (VPiT)—a simple extension to visual instruction tuning which builds upon the existing paradigm of passing continuous visual tokens as input to the LLM. VPiT trains an LLM to output *both continuous visual tokens and discrete text tokens* in the finetuning stage. The model takes pretrained vision encoder embeddings as well as text tokens as input, and outputs a combination of text tokens and continuous visual tokens. To visualize the generated visual tokens, we finetune a diffusion model to map the embeddings back into pixel space (see Fig. 1 for an example). This framework allows us to study the synergy between visual understanding, visual generation, and pretrained LLMs, which leads to several intriguing findings outlined below.

First, we show that the ability to predict visual tokens emerges from understanding visual inputs and requires minimal additional training. Similar to visual instruction tuning,

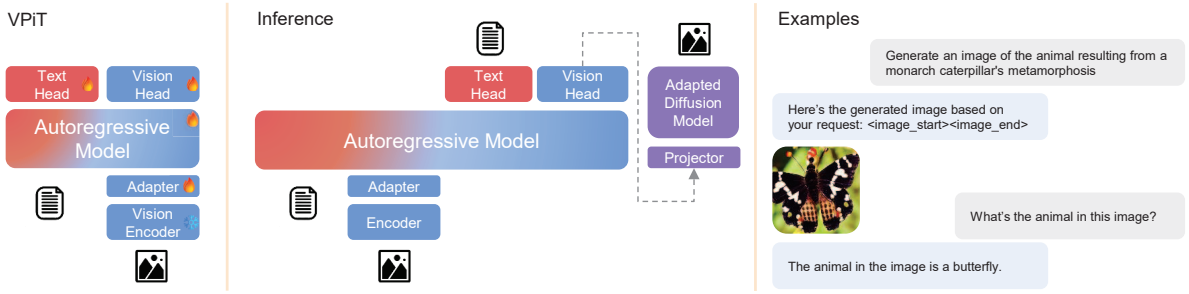


Figure 1. **VPiT Training, Inference, and Examples of MetaMorph. Left**: In Visual-Predictive Instruction Tuning (VPiT), we finetune a pretrained LLM to generate both text and visual tokens using separate text and vision heads. **Middle**: During inference, the model accepts an arbitrary input sequence of image(s) and text and outputs discrete text tokens and continuous visual tokens. These visual tokens can be visualized via a separately finetuned diffusion model, which is trained to condition on the vision encoder's output. **Right**: An example conversation from MetaMorph trained with VPiT. Here, the model implicitly solves a visual puzzle to generate a butterfly. The conversation continues with new user questions as the model continues to process autoregressively, independent of the diffusion-based visualization.

VPiT efficiently and effectively *morphs* an LLM into an "unified" model that understands and generates multimodal tokens. When trained jointly with sufficient visual understanding data, this process requires as little as *200k* additional visual generation data.

We further establish that the abilities to understand and generate visual tokens are *intrinsically linked* and *asymmetrical*. Specifically, increasing understanding data improves visual understanding (measured by higher VQA scores) and generation (measured by lower FID scores). Conversely, increasing generation data enhances generation quality and also contributes to stronger visual understanding—but to a lesser degree. Importantly, our findings highlight an asymmetry in how training each ability impacts the model's overall vision performance: understanding-centric training substantially outperforms generation-centric training in improving both visual understanding and generation.

Building upon these findings, we train a unified model called *MetaMorph* to predict multimodal tokens with VPiT. We leverage diverse data sources ranging from common visual question answering datasets to pure image and video data without text annotations. MetaMorph achieves competitive performance on both visual understanding and visual generation benchmarks. Furthermore, we show this unified modeling approach allows models to leverage the power of LLMs. For instance, MetaMorph can extract knowledge from the pretrained LLM when generating visual tokens. More surprisingly, we observe that MetaMorph can implicitly perform reasoning steps before generating visual tokens—e.g. when prompted with *"the animal resulting from a monarch caterpillar's metamorphosis"*, MetaMorph successfully generates an image of a butterfly (Fig. 1).

Our results suggest that 1) training a unified model with instruction tuning is feasible, and 2) LLMs have strong pre-existing visual capabilities which can be activated using significantly fewer samples compared to extensive pretraining. These insights shed light on the development of mixed-modality models. As the community continues to improve visual understanding in Multimodal LLMs [32, 63, 67] by advancing base LLMs, instruction tuning techniques, and data, we highlight that these efforts may also implicitly lead to models that are better at visual generation.

2. Visual-Predictive Instruction Tuning

Visual instruction tuning as introduced by LLaVA [36] demonstrates that LLMs can be taught to understand visual inputs. This is achieved by finetuning on million-scale data. The success of late-fusion instruction tuning suggests that LLMs may already possess innate visual understanding ability. This ability simply needs to be unlocked through lightweight finetuning. Analogously, we hypothesize that LLMs already possess a degree of innate visual generation ability which just needs to be unlocked with finetuning.

Motivated by this, we present Visual-Predictive Instruction Tuning (VPiT, Fig. 1)—a simple design which extends existing instruction tuning methods to additionally generate visual tokens rather than text alone. We use the same architecture and next-token prediction paradigm to unlock visual generation capabilities without bells and whistles. We take a pretrained LLM and finetune it to predict both discrete text tokens and continuous visual tokens. The visual tokens can be visualized with an adapted diffusion model.

2.1. From Unimodal to Multimodal Next-Token Prediction

The standard instruction tuning setup consists of an input sequence of conversation rounds [61, 69]: $(P_i, R_i)_{i=1}^{N}$, where P_i and R_i represent prompts and responses for the i-th round of conversation, respectively. The model is trained to generate responses based on the prompt. VPiT adds the following mechanisms to a standard instruction tuning setup to unlock visual understanding and generation.

Tokenizing multimodal data. We extend P_i and R_i to include both text and images. To integrate visual data into a pretrained LLM, we process data closely following visual instruction tuning [36]:

- **Text Data**: Text is tokenized into discrete tokens with a standard tokenizer used by the LLM.
- **Visual Data**: Images are encoded with a pretrained vision encoder such as SigLIP [81]. The output is continuous visual tokens which are then interpolated to $m = 64$ tokens. To pass the visual tokens as input to the LLM, we apply a trainable projection layer to align the dimensions.

Model architecture. We take a pretrained LLM and finetune it to process arbitrary sequences of text and visual tokens (detailed next in Sec. 2.2). We keep the original LLM head for text prediction, and attach a separate vision head to the LLM for predicting visual tokens, i.e., the output tokens generated by the vision encoder when processing images. The vision head is a projection layer that projects from the LLM's dimension to the vision encoder's dimension. All response tokens can then be trained and predicted autoregressively, with prompt tokens as context.

Unlike conventional visual instruction tuning, in VPiT, visual tokens are also outputs of the LLM—not just inputs. To make the LLM aware of the presence of visual tokens, we introduce special tokens ⟨image_start⟩ and ⟨image_end⟩ to indicate the boundaries of visual token sequences and when to use the vision head.

Loss functions. The language head outputs a probability distribution over the vocabulary and is trained with cross-entropy loss for next-token prediction. Visual prediction uses cosine similarity loss between the LLM's predicted visual tokens and those from the vision encoder. Consistent with instruction tuning practices, the model only makes predictions and incurs loss on response tokens.

2.2. Using Broad Types of Data

Because VPiT enables the model to predict both text and visual tokens in its responses, it allows the use of a broader range of training data. Traditional visual instruction tuning, on the other hand, primarily relies on question-and-answer pairs. The majority of our dataset is publicly available, and we categorize it into three major categories below. This categorization enables us to systematically study the model, as detailed in Sec. 3 and Sec. 4. All data types are formatted as instruction tuning style prompt & response pairs. See further details in Appendix C.2.

1. **Visual Understanding Data**: This category includes data that takes image(s) or video as input and outputs text responses. See Fig. 1 for an example. We use:
   - **ImageQA:** Cambrian-7M [63]. The model answers questions based on input image(s).
     $P_i \in \{\langle\texttt{visual tokens}\rangle, \langle\texttt{text prompt}\rangle\}$
     $R_i \in \{\langle\texttt{text response}\rangle\}$
   - **VideoQA:** VideoStar [87] and ShareVideo [83]. The model answers questions based on the input video. For videos in VideoQA, we process frames at 1 FPS.
     $P_i \in \{\langle\texttt{visual tokens}\rangle, \cdots, \langle\texttt{visual tokens}\rangle, \langle\texttt{text prompt}\rangle\}$
     $R_i \in \{\langle\texttt{text response}\rangle\}$
2. **Visual Generation Data**: MetaCLIP [76]. The model predicts visual tokens based on an image description. We using at most 5 million pairs. We curate the data into question-answering formats.
   $P_i \in \{\langle\texttt{text prompt}\rangle\}$
   $R_i \in \{\langle\texttt{text response}\rangle, \langle\texttt{visual tokens}\rangle\}$
   We prompt the model to generate visual tokens with instructions like *"Generate an image of..."*. The text responses are *"Here is an image based on your request..."*. See Fig. 1 for an example.
3. **Other Visual Data**: This category includes data that requires the model to predict visual tokens given interleaved input visual tokens and text tokens. We use:
   - **Video Data:** SomethingSomethingV2 [21] and HowTo100M [47]. The model predicts frames in a sequential order. We design different question-answer pairs to probe into the video, such as asking about future frames, past frames, and reordering frames.
     $P_i \in \{\langle\texttt{visual tokens}\rangle, \cdots, \langle\texttt{visual tokens}\rangle, \langle\texttt{text prompt}\rangle\}$
     $R_i \in \{\langle\texttt{visual tokens}\rangle, \cdots, \langle\texttt{visual tokens}\rangle\}$
   - **Visual Thinking Data:** Visualization-of-Thought [57] and VStar [72]. The model predicts multimodal tokens in its response before addressing problems. For instance, it predicts a zoomed-in view of an image before generating textual responses.
     $P_i \in \{\langle\texttt{visual tokens}\rangle, \langle\texttt{text prompt}\rangle\}$
     $R_i \in \{\langle\texttt{text response}\rangle, \langle\texttt{visual tokens}\rangle, \langle\texttt{text response}\rangle\}$
   - **Image-to-Image Data:** InstructPix2Pix [10] and Aurora [28]. The model generates a transformed image conditioned on a text description and an input image.
     $P_i \in \{\langle\texttt{visual tokens}\rangle, \langle\texttt{text prompt}\rangle\}$
     $R_i \in \{\langle\texttt{visual tokens}\rangle\}$

2.3. Mapping Tokens to Images through Diffusion

Because models trained with VPiT learn to predict continuous visual tokens, we need to map the predicted tokens back into pixel space. We leverage the concept of a "Diffusion Autoencoder" [9, 27, 34, 50, 51] in which the diffusion model can be adapted to condition on image embeddings. Specifically, we finetune an existing diffusion model to condition on outputs from the vision encoder using held-out data.

At inference time, if ⟨image_start⟩ is generated, the model begins outputting visual tokens until ⟨image_end⟩. We then plug the generated visual tokens into the diffusion model to visualize the prediction in pixel space. We use latent diffusion model training procedures. Details on the hyperparameters and training are provided in Appendix A.2.

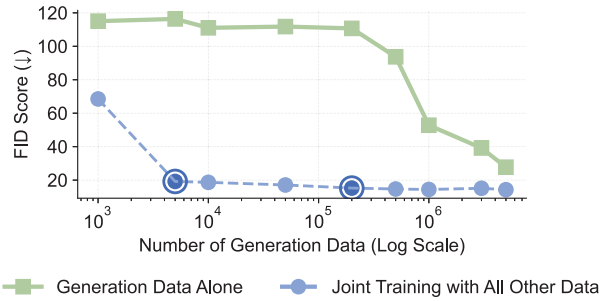


Figure 2. **Generation-only training vs. Joint training with other data.** Training solely on generation data results in inferior performance. Joint training with additional data enables visual generation with only 5k generation data and yields high-quality outputs with 200k generation data.

3. Findings on Unlocking Visual Generation

We study the following questions about the effects and synergy of visual understanding and generation, under VPiT:
§3.1 Can visual generation be unlocked through lightweight tuning, or does it require extensive data?
§3.2 Are visual understanding and generation mutually beneficial or orthogonal?
§3.3 How much does more visual understanding or generation data contribute to understanding and generation quality?
§3.4 Which visual understanding tasks correlate the most with generation performance?

Evaluation settings. We use 9 ImageQA benchmarks (MM-Bench, Seed, VStar, MMVP, MMMU, ChartQA, TextVQA, ScienceQA, RealWorldQA) to evaluate different aspects of the model. For image generation, we use the finetuned diffusion model to visualize generated visual tokens and measure FID score (lower is better) and CLIP score (higher is better) on the COCO-30K dataset. Unless otherwise specified, we use LLaMA-3 8B [3] / SigLIP ViT-SO400M-14@384 [81] as the pretrained LLM / vision encoder. We also study the effect of different LLMs in Sec. 3.2. We use instruction tuned versions of the LLMs. We pretrain the adapter between the vision encoder and the LLM following visual instruction tuning. For experiments in this section, we provide training details in Appendix A and full results in Appendix B.

3.1. Visual Generation Can Be Unlocked Efficiently by Joint Training with Visual Understanding

We start by investigating the number of image-text samples required to teach a language model to generate high-quality visual tokens. To this end, we randomly sample {1k, 5k, 10k, 50k, 200k, 1M, 3M, 5M} image-text pairs from our generation data (MetaCLIP dataset [76]). We explore two settings: (1) finetuning the LLM using only visual generation data, and (2) joint training visual generation with visual understanding and the rest of data types described in Sec. 2.2.

In Fig. 2, we see that training solely on visual generation performs significantly worse than joint training with all other data. With over 3 million image-text pairs, the model struggles to generate high-quality visual images ($\sim$ 40 FID score), and performance remains inferior to joint training with 5 million pairs. This suggests that training solely on visual generation data is significantly less sample efficient. This finding aligns with a prior study [84] which also suggests that LLMs cannot be easily tuned to generate visual tokens when trained with only generation data. In contrast, joint training with other datasets substantially improves generation performance. The model generates effective visual tokens with just *5k* generation data, and performance stabilizes around *200k* samples. This indicates that visual generation is not an orthogonal capability but rather an ability that emerges more effectively with joint training.

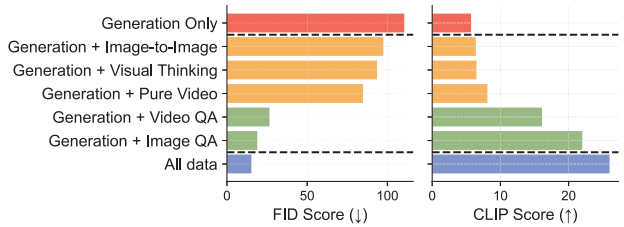


Figure 3. **Impact of different data types on visual generation.** The baseline of training on only visual generation data is red; Joint training with other data is yellow; Joint training with visual understanding data is green; and all data is blue. Joint training with additional data improves the baseline, with visual understanding tasks contributing the most to enhancing visual generation.

To better understand how each type of data contributes to visual generation, we conduct a controlled experiment using 200k visual generation data, joint training individually with each data type defined in Sec. 2.2. We also compare them with training all the data together and show results in Fig. 3. While all data types enhance the model's visual generation, the degree of improvement varies. Visual understanding data, such as ImageQA and VideoQA, significantly boost the model's visual generation capabilities. This indicates a strong link between the ability to understand visual content and generate visual tokens. Additionally, combining all data types in training further improves performance, suggesting that the benefits from different data types can be additive.

> ***Finding 1:*** Visual generation can be unlocked with significantly less generation data when the model is jointly trained with visual understanding data, in contrast to training only on generation data.

3.2. Visual Understanding and Generation are Mutually Beneficial

More understanding data leads to better understanding and generation. Building upon findings from the previous subsection, we perform a controlled experiment to investigate how visual understanding correlates with visual genera-

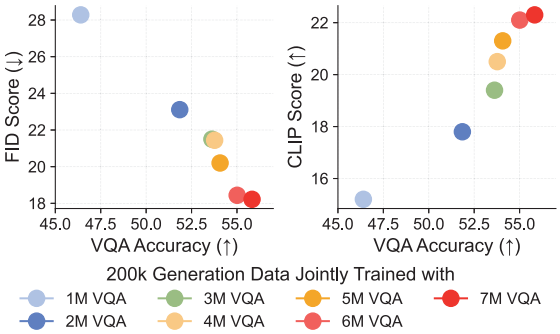


Figure 4. **VQA Performance vs. Generation Performance with generation data controlled at 200k.** Increasing understanding data improves VQA and generation performance.

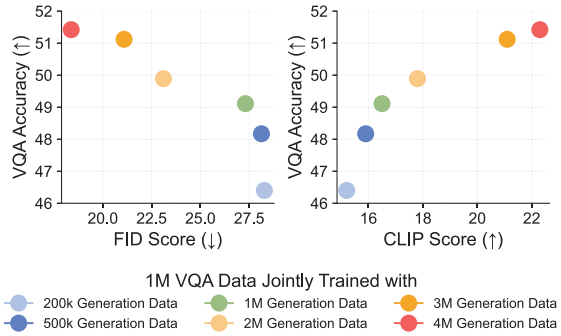


Figure 5. **Generation Performance vs. VQA Performance with VQA data controlled at 1M.** Increasing generation data improves generation and VQA performance.

tion ability. We ablate our model using a fixed set of 200k generation data while varying VQA data from 1M to 7M samples from Cambrian-7M to develop different levels of visual understanding. The results presented in Fig. 4 indicate that stronger VQA ability correlates with better generation.

More generation data leads to better understanding and generation. Here, we investigate the reverse direction: does enhancing the model's visual generation capability also relate to higher VQA performance? To explore this, we conduct a controlled experiment using 1M fixed VQA samples as the baseline for understanding. We then vary the amount of generation data ({200k, 500k, 1M, 2M, 3M, 4M}) to adjust generation capacity while joint training with the fixed 1M VQA data. We present results in Fig. 5. Within the 1M VQA setting, stronger generation ability is correlated with improved VQA performance. This implies that increasing the amount of generation data not only enhances generation but also positively impacts VQA performance.

This synergy scales across different LLMs. We examine whether the findings transfer across LLM backbones. Using a data composition of 7M VQA and 1M generation data, we train VPiT on LLaMA-3 8B, LLaMA-3.1 8B, and LLaMA-3 70B. Fig. 6 shows scaling behaviors across different LLMs.

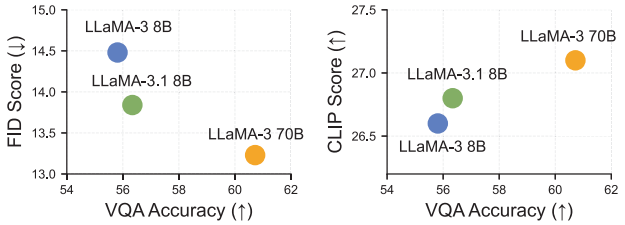


Figure 6. **Comparison between different language backbones.** We jointly train 7M VQA and 1M Generation data on different language backbones (LLaMA-3 8B, LLaMA-3.1 8B, LLaMA-3 70B). We observe that the synergy between understanding and generation transfer across LLMs.

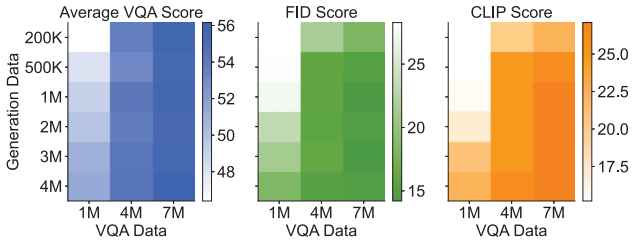


Figure 7. **Heatmap visualization of Average VQA Score, FID Score, and CLIP Score across varying amounts of VQA data and generation data.** Darker colors indicate better performance. Increasing VQA data is more effective for improving both understanding and generation capabilities.

> ***Finding 2:*** Visual understanding and generation are synergistic. Increasing data for either capability enhances both simultaneously.

3.3. Understanding Data Contributes More

We investigate whether understanding and generation data contribute equally. Here, we jointly train different scales of VQA data {1M, 4M, 7M} and generation data {200k, 500k, 1M, 2M, 3M, 4M}. Fig. 7 summarizes these findings, with the x-axis representing VQA data, and the y-axis representing generation data. Results are visualized on heatmaps using darker colors for better performance.

The results indicate that increasing VQA data yields the most significant improvements in all three metrics. When VQA data is relatively low (1M), increases in generation data lead to noticeable improvements, as reflected by the gradual darkening in the plot. However, as the VQA data scales up (from 1M to 4M to 7M), the impact of VQA data becomes more pronounced, demonstrated by a sharp color transition in the heatmap. Ultimately, with 7M VQA data, increases in generation data contribute minimally. These results demonstrate the critical role of understanding data in enhancing both understanding and generation performance.

> ***Finding 3:*** While increasing data improves performance overall, the impact of visual understanding data is more significant than visual generation data.

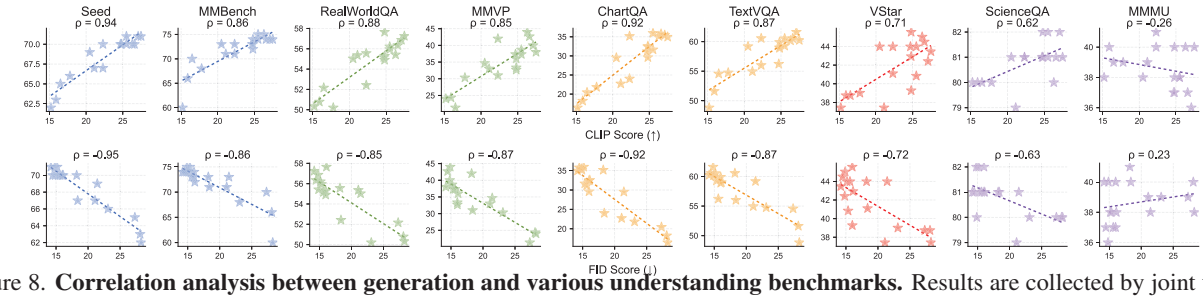


Figure 8. **Correlation analysis between generation and various understanding benchmarks.** Results are collected by joint training different amounts of VQA data combined with varying quantities of generation data. Each subplot shows the correlation (ρ) with a fitted regression line. Stars represent data points. We analyze General VQA, Vision-Centric VQA, Text&Chart VQA, High-Resolution VQA, and Knowledge VQA. For most tasks, generation performance and VQA performance are strongly correlated: higher VQA performance indicates better generation and vice versa. Knowledge-intensive and high-resolution VQA tasks exhibit weaker correlations with generation.

3.4. Certain Understanding Tasks Correlate More with Generation Performance

Given the diverse nature of understanding tasks such as OCR, Vision-Centric tasks, and Knowledge-based tasks, we investigate which tasks most strongly correlate with generation ability. Inspired by Cambrian-1, we categorize VQA tasks into five groups: General, Text&Chart, High-Resolution, Knowledge, and Vision-Centric VQA. Using the results from our earlier experiments, which jointly train various VQA data scales with different amounts of generation data, we plot each benchmark's VQA performance against generation performance in Fig. 8. We also calculate the Pearson correlation (ρ) between VQA scores and FID/CLIP Scores.

Fig. 8 shows that General, Vision-Centric, and Text&Chart VQA tasks strongly correlate with generation performance, each with a ρ above 0.85. High-Resolution VQA exhibits moderate correlation, with ρ around 0.7. In contrast, Knowledge VQA tasks show weak correlation with generation performance. These findings suggest that generation ability aligns more closely with the model's vision capabilities rather than knowledge-specific tasks.

> ***Finding 4:*** General, vision-centric, and text VQA tasks exhibit strong correlations with visual generation, whereas knowledge VQA tasks do not.

4. MetaMorph Model

Based on the insights in Sec. 3, we train our unified model, MetaMorph, based on LLaMA-3.1 8B [3], using VPiT with the data in Sec. 2.2. We present our results in three parts: quantitative performance (Sec. 4.1), evidence of MetaMorph leveraging LLM knowledge in visual generation (Sec. 4.2), and multimodal implicit reasoning skills. (Sec. 4.3).

4.1. Competitive Performance

We compare MetaMorph with other unified models and summarize results in Tab. 1. Since these models are trained on different datasets and base LLMs (or pretrained from scratch), an *apples-to-apples* comparison is difficult. Nevertheless, MetaMorph demonstrates competitive performance and outperforms other unified models on most benchmarks—even when prior models may have been trained on more data. Compared to models trained from scratch, such as EMU-3 [68] and Chameleon [62], MetaMorph leverages latest LLMs and achieves competitive understanding and generation performance. MetaMorph highlights that unified models can be developed effectively from pretrained LLMs.

4.2. MetaMorph can Leverage LLM Knowledge

MetaMorph effectively leverages the world knowledge embedded in pre-trained LLMs. We show examples on the left side of Fig. 9. We prompt the model to generate concepts requiring non-trivial and specialized knowledge. Examples include *"Chhogori"* (the world's second-highest mountain), *"Oncilla"* (a small wildcat from South America), and *"Chizarira"* (an isolated wilderness area in Zimbabwe).

MetaMorph successfully translates domain-specific knowledge into accurate visual tokens, thereby displaying the ability to *leverage world knowledge from LLMs*. In contrast, the *latest* Text-to-Image (T2I) model, Stable Diffusion-3.5 8B, struggles to generate the correct concept despite producing high-quality images. This issue may stem from the text embedding models it uses——CLIP [52] and T5 [54]——which fail to properly encode these specialized terms [80].

On the right side of Fig. 9, we demonstrate how MetaMorph handles common semantic challenges more effectively than text embedding models such as CLIP and T5. These challenges include negation and subjectivity [64]. MetaMorph differentiates semantic nuances such as "slightly" versus "very" and "few" versus "many", which are common failures in existing text-to-image systems.

4.3. Reasoning in Multimodal Generation

In Fig. 10, we present examples where the model generates images in response to puzzle prompts such as *"The national flag of the country where Yellowstone National Park is located"*. For each puzzle, we directly use the prompt *"Generate an image of {puzzle}"*, without calling Chain-of-Thought

Method	Base LLM	Image QA									Video QA	Generation
		MMBench^{EN}	SEED	RealworldQA	MMVP	SQA	MMMU	VStar	ChartQA	TextVQA	MV-Bench	COCO (FID)
Visual Understanding Only												
GPT-4V*		75.8	69.1	61.4	50.0	75.7	56.8	55.0	78.5	78.0	43.5	-
Visual Generation Only												
Stable Diffusion 1.5*		-	-	-	-	-	-	-	-	-	-	9.6
Dalle 2*		-	-	-	-	-	-	-	-	-	-	10.4
Imagen*		-	-	-	-	-	-	-	-	-	-	7.3
Unified Models												
EMU-3*		58.5	68.2	57.4	36.6†	89.2	31.6	51.8†	68.6	64.7	-	12.8
Janus*	DeepSeek 1.3B	69.4	63.7	-	-	-	30.5	-	-	-	-	8.5
VILA-U$_{256}$†	LLaMA-2 7B	66.6	57.1	46.6	22.0	67.1	32.2	38.7	11.4	48.3*	40.8	19.6
Transfusion*		-	-	-	-	-	-	-	-	-	-	6.7
Chameleon-7B†		35.7	27.2	19.6	0.0	50.3	28.4	37.1	0.0	0.0	-	26.7*
MetaMorph (ours)	LLaMA-3.1 8B	75.2	71.8	58.3	48.3	83.2	41.8	44.0	37.1	60.5	48.8	11.8

Table 1. **Comparison of MetaMorph with other unified models.** MetaMorph offers competitive performance compared to other leading unified models. Models in gray are understanding-only or generation-only. Unified models without a base LLM are trained from scratch. *We use numbers reported in original papers. †We obtain results using official open-sourced model weights.

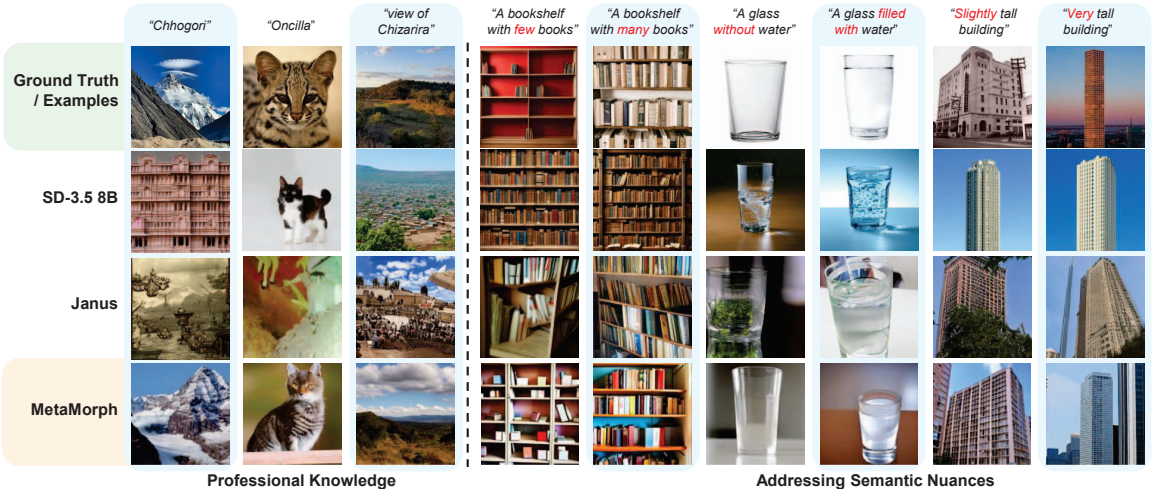


Figure 9. **Examples of MetaMorph leveraging LLMs to generate visual tokens. Left**: MetaMorph can leverage knowledge from the LLM to generate visual tokens for professional terms that need domain-specific understanding. **Right**: MetaMorph also avoids common mistakes seen in T2I models that condition on text embeddings (e.g., Stable Diffusion-3.5 8B).

(CoT) [70] in the prompts. MetaMorph generates correct images from prompts that require multi-step reasoning.

For example, when answering the question *"A musical instrument, this instrument is often played by the scientist who formulated the theory of special relativity"*, the model needs to implicitly complete three reasoning steps: it identifies Albert Einstein as the scientist formulated the theory of special relativity, recognizes that his preferred instrument is the violin, and then directly generates correct visual tokens—a violin—without explicitly separating these steps during the generation process. This result implies that MetaMorph implicitly solves the puzzle and generates correct visual tokens immediately following the prompt. These results align with the findings in *Physics of LMs* [5, 77], that LLMs precompute reasoning graphs before autoregressively generating subsequent tokens. Here, we demonstrate that this capability transfers to the unified multimodal model setting.

5. Related Work

Instruction tuning and visual instruction tuning. Instruction tuning [61, 69] finetunes a pretrained LLM to learn the

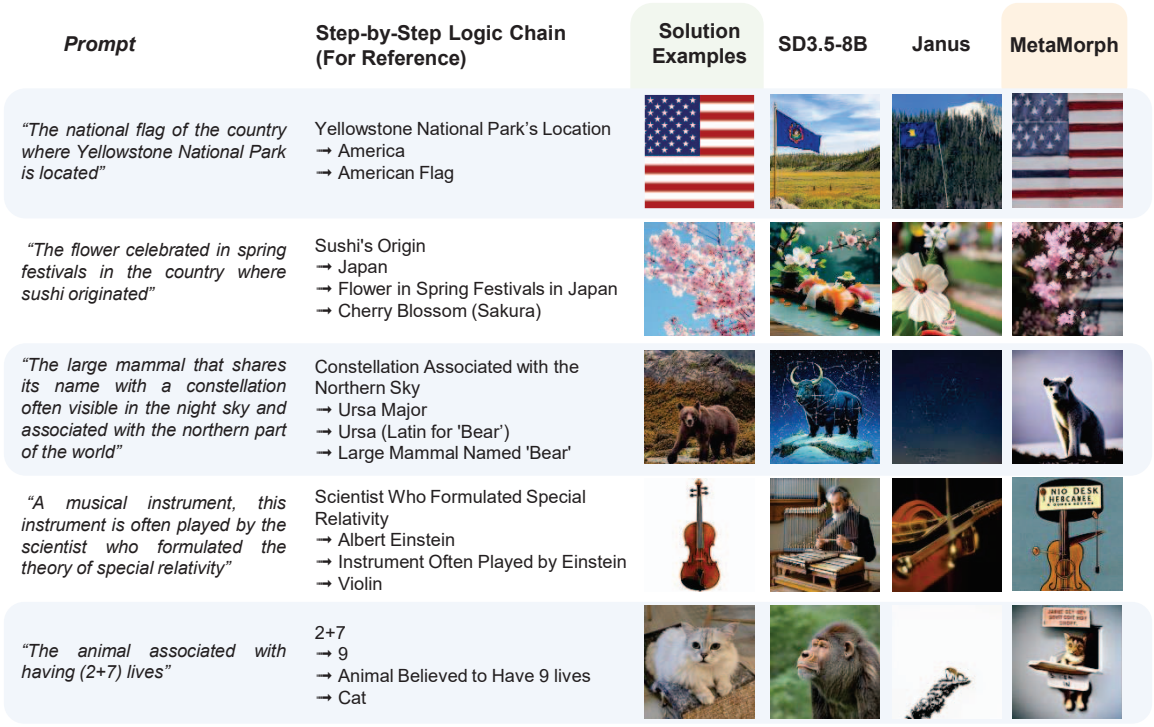


Figure 10. **Examples of MetaMorph solving reasoning problems in visual generation.** We design puzzles that require multi-step reasoning. We include reference **logic chains** needed to solve the puzzles, and reference **solution examples**. When prompting each model, we directly feed in the puzzle without any CoT hints or logic chains. MetaMorph has the ability to implicitly solve these puzzles and generate the correct image without explicitly creating or processing a logic chain.

format and unleash the capabilities acquired during pretraining [85]. LLaVA [36] extends instruction tuning into the multimodal domain. Different lines of work focus on improving data [11, 29, 30], visual representation [13, 26, 63], and training strategies [18, 38]. Using only a few million data, open-source MLLMs [32, 38, 63] reach performance levels comparable to proprietary models [6, 48] on a number of benchmarks [40, 78, 79] and applications [49, 82].

From Multimodal LLMs to unified models. Recent efforts to construct unified models have primarily relied on either extensive pretraining or heavy fine-tuning on billion-scale datasets. Some studies also use continuous embeddings for predicting visual tokens, integrating regression losses [59, 60], or leveraging diffusion methods [15]. Other approaches [1, 39, 42, 43, 62, 68, 73] tokenize all data into discrete tokens. Recent research also explored combining autoregressive and diffusion objectives [75, 86]. Different from previous study, we demonstrate that unified models can be effectively trained in low-data regimes during instruction tuning, while also providing insights into the reciprocal relationship between visual understanding and visual generation.

6. Discussion

In this work, we propose VPiT—a simple yet effective extension to visual instruction tuning—that enables LLMs to predict multimodal tokens. VPiT unlocks the use of a more diverse range of instruction tuning data than just visual question answering, such as text-to-image data. Through controlled experiments, we find that visual generation ability emerges as a natural byproduct of improved visual understanding and requires modest additional generation data. In addition, we find that while visual understanding and generation are mutually beneficial, adding more visual understanding data disproportionately improves overall performance compared to adding more generation data.

Leveraging these insights, we train MetaMorph by fine-tuning LLaMA-3.1 8B with VPiT. With a simple training process, MetaMorph achieves competitive performance in both visual understanding and generation. Qualitative evaluation of our model shows that MetaMorph can leverage world knowledge and reasoning abilities of the base LLM during visual generation. For example, it can perform multimodal tasks that typically require multiple-step reasoning, such as generating images of specialized proper nouns (*"Chhogori"*) or solving visual puzzles (*"generate an image of the animal resulting from a monarch caterpillar's metamorphosis"*). This indicates that LLMs already possess a degree of "prior" visual knowledge which can be activated with only minimal instruction tuning with VPiT. Overall, LLMs may have a similar representation space as unified and multi-functional models [25]. We hope these insights inspire more exploration toward developing models for general intelligence.

7. Acknowledgements

ST is supported by the Meta AI Mentorship (AIM) Program. SX also acknowledges support from the OpenPath AI Foundation, IITP grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project) and NSF Award IIS-2443404.

References

[1] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022. 8

[2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 1

[3] AI@Meta. Llama 3 model card. 2024. 1, 4, 6

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[5] Zeyuan Allen-Zhu. ICML 2024 Tutorial: Physics of Language Models, 2024. Project page: https://physics.allen-zhu.com/. 7

[6] Anthropic. Claude, 2024. 8

[7] Jimmy Lei Ba, Jamie Kiros, and Geoffrey E. Hinton. Layer normalization. In *NeurIPS*, 2016. 13

[8] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. In *TMLR*, 2024. 13

[9] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. In *TMLR*, 2022. 3

[10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3, 14, 15

[11] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 8

[12] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024. 14

[13] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 8

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2024. 1

[15] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024. 8

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[17] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2024. 1

[18] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 8

[19] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 13

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 14, 15

[21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 3

[22] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 13

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 13

[24] 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 *NeurIPS*, 2017. 13

[25] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024. 8

[26] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *ECCV*, 2025. 8

[27] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. In *NeurIPS*, 2024. 3

[28] Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulations. In *NeurIPS*, 2024. 3, 14, 15

[29] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[30] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 8

[31] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 13

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 8

[33] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 13

[34] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *NeurIPS*, 2024. 3

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 16

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 8

[37] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1

[38] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 8

[39] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 1, 8

[40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024. 8, 13

[41] I Loshchilov. Decoupled weight decay regularization. In *ICLR*, 2019. 13

[42] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2022. 8

[43] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, 2024. 8

[44] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 13

[45] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 13

[46] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 13

[47] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 3, 14, 15

[48] OpenAI. gpt4o, 2024. 8

[49] Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. Autonomous evaluation and refinement of digital agents. In *COLM*, 2024. 8

[50] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *ICLR*, 2024. 3

[51] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 3

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 6

[53] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 13

[54] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2019. 6

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 13

[56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1, 13, 14

[57] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *NeurIPS*, 2024. 3, 14, 15

[58] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. 13

[59] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 8

[60] 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. In *ICLR*, 2024. 8

[61] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B.

Hashimoto. Alpaca: A strong, replicable instruction-following model, 2023. 2, 7

[62] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 6, 8

[63] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 1, 2, 3, 8, 13, 14, 16

[64] Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. In *NeurIPS*, 2024. 6

[65] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024. 13

[66] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. 2023. 1

[67] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2

[68] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 6, 8

[69] 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 *ICLR*, 2022. 2, 7

[70] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 7

[71] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 1

[72] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024. 3, 13, 14, 15

[73] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 1, 8

[74] xAI. grok, 2024. 13

[75] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1, 8

[76] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2024. 3, 4, 13, 14

[77] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process. *ArXiv e-prints*, abs/2407.20311, 2024. Full version available at http://arxiv.org/abs/2407.20311. 7

[78] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 8, 13

[79] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 8

[80] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022. 6

[81] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 3, 4

[82] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *NeurIPS*, 2024. 8

[83] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 3, 14

[84] Yuhui Zhang, Brandon McKinzie, Zhe Gan, Vaishaal Shankar, and Alexander Toshev. Pre-trained language models do not help auto-regressive text-to-image generation. In *EMNLP*, 2023. 4

[85] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. In *NeurIPS*, 2024. 1, 8

[86] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 1, 8

[87] Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-levy. Video-star: Self-training enables video instruction tuning with any supervision. In *arXiv preprint arXiv:2407.06189*, 2024. 3, 14