UniToken: Harmonizing Multimodal Understanding and Generation through Unified Visual Encoding

Yang Jiao[1,2,3*], Haibo Qiu[3*], Zequn Jie[3], Shaoxiang Chen[3], Jingjing Chen[1,2],
Lin Ma[3], and Yu-Gang Jiang[1,2]

[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3] Meituan

yjiao23@m.fudan.edu.cn, haibo-qiu@outlook.com
{zequn.nus, forest.linma}@gmail.com, {sxchen13, chenjingjing, ygj}@fudan.edu.cn

Abstract

We introduce UniToken, an auto-regressive generation model that encodes visual inputs through a combination of discrete and continuous representations, enabling seamless integration of unified visual understanding and image generation tasks. Unlike previous approaches that rely on unilateral visual representations, our unified visual encoding framework captures both high-level semantics and low-level details, delivering multidimensional information that empowers heterogeneous tasks to selectively assimilate domain-specific knowledge based on their inherent characteristics. Through in-depth experiments, we uncover key principles for developing a unified model capable of both visual understanding and image generation. Extensive evaluations across a diverse range of prominent benchmarks demonstrate that UniToken achieves state-of-the-art performance, surpassing existing approaches. These results establish UniToken as a robust foundation for future research in this domain. The code and models will be available at https://github.com/SxJyJay/UniToken.

1. Introduction

With next-token prediction demonstrating promising scaling laws in both language modeling [51, 2, 3] and multimodal comprehension [35, 4, 24], exploring similar scaling potential in the field of visual generation [50, 53] has become a recent trend. LlamaGen [50] and OmniTokenizer [53] demonstrate the scalability of transformer architectures in the domains of image and video generation, respectively. Sharing similar model architectures and learning paradigms, the unification of multimodal understanding and generation within a shared model offers promising prospects and has recently garnered significant research interest [41, 54, 56, 65, 55]. Existing works can be categorized into two groups based on their visual encoding paradigms. The first group, exemplified by Chameleon [41] and Emu3 [54], encodes visual inputs into discrete tokens predefined within a fixed vocabulary. The second group, represented by Unified-IO2 [39] and Janus [55], adopts a decoupled visual encoding approach, utilizing continuous visual tokens for image comprehension and discrete tokens for image generation.

However, these two paradigms face challenges stemming from their respective visual encoding techniques. For the discrete-only encoding paradigm, the efficiency of absorbing knowledge from multimodal data lags significantly behind that of its continuous-only counterpart, as proved in Tab.5. This phenomenon arises from the information loss incurred during the quantization process used to generate discrete tokens. Moreover, discarding the continuous visual tokens sacrifices the sweetness of advanced image encoding techniques [33, 42]. On the other hand, in the decoupled encoding paradigm, employing different visual encoders for distinct tasks restricts the model's flexibility, and the burdens introduced by such mode switching increase as the model's functionality continues to scale up.

Toward this end, in this paper, we propose UniToken, which employs a unified visual encoding that is agnostic to specific tasks. As shown in Fig.1, we encode an image with dual visual encoders, resulting in a combination of discrete and continuous tokens that form the unified visual representation. Carrying both low-level details and high-level semantics, this unified visual representation can seamlessly support both image understanding and generation tasks. Additionally, we employ several advanced tech-
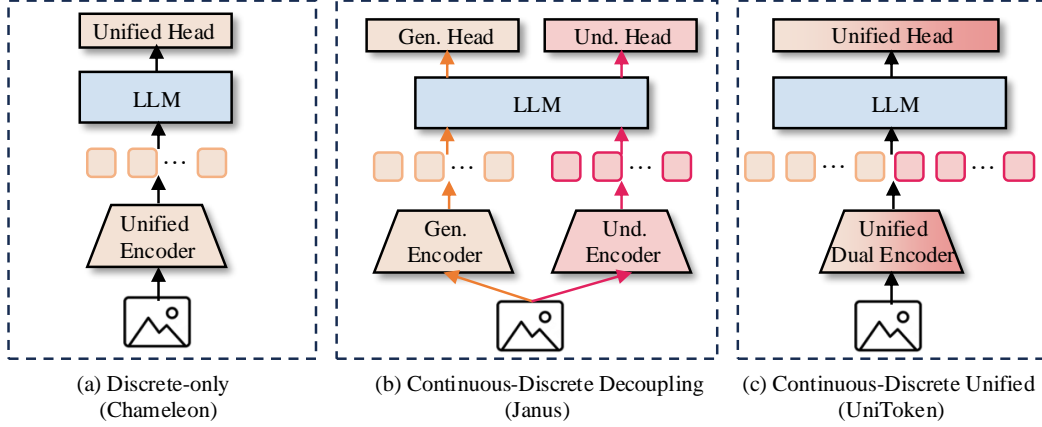

Figure 1. **Different visual encoding paradigms for developing a unified model for visual understanding and image generation.** We use orange to denote components related to discrete visual encoding and red to signify components associated with continuous visual encoding. For the sake of brevity, we omit the text tokens in the image.

niques, namely scaling up resolutions and tuning the ViT, to significantly amplify the visual representation. By incorporating these techniques, multimodal comprehension capabilities, especially in text-reading-related scenarios, are significantly boosted.

Through extensive experiments, we uncover several valuable empirical insights for developing a unified multimodal understanding and generation model: (1) **Task interference:** a combined discrete and continuous visual encoding ensures seamless compatibility between multimodal understanding and generation tasks, while relying solely on discrete visual tokens is prone to task interference. (2) **Task-specific data distribution:** the data proportion between multimodal understanding and generation tasks should be adjusted when using training data of varying scales. Furthermore, through systematic evaluation across a wide array of prevalent visual understanding and image generation benchmarks, our UniToken outperforms unified models for both understanding and generation and competes with state-of-the-art specialists in each respective field.

In summary, our contributions are three-fold:

- We propose UniToken, which integrates both discrete and continuous tokens as a unified visual representation to deal with diverse tasks.

- We further equip our UniToken with advanced visual encoding techniques, significantly boosting our model's comprehension capabilities.

- We uncover several empirical insights of developing a unified model for multimodal understanding and generation. And our UniToken can compete with state-of-the-art approaches in both multimodal understanding and generation fields.

2. Related Works

2.1. Multimodal Understanding

Building upon powerful Large Language Models (LLMs), MLLMs demonstrate impressive multimodal content comprehension and reasoning capabilities. To efficiently adapt visual concepts to the textual world of LLMs, mainstream MLLMs [35, 4, 8, 24] leverage vision encoders such as CLIP [46] or SigLIP [63], which are enriched with semantics derived from vision-text contrastive training. By further tuning an adapter and the entire model sequentially with multimodal instructional data, knowledge encapsulated in pretrained LLMs can effectively fuel the comprehension of the visual world. However, the visual features from these vision encoders inherently lack the image generation potential. To mitigate this gap, recent advancements [19, 20, 25, 14] integrate an external diffusion model with the MLLM, leveraging MLLM outputs to generate the conditions for the diffusion process. The generative capability of this paradigm may be limited by the external diffusion model and the effectiveness of adapting the MLLM's outputs to the diffusion conditional space.

2.2. Visual Generation

To explore the scaling property in the visual generation field, attention has been switched to the autoregressive-based paradigm. VQ-VAE [52] and VQ-GAN [16] are pioneers of this paradigm, where the image is first quantized into a sequence of discrete codebook IDs, and then their underlying distribution is modeled with transformer architectures. To reduce the quantization error, recent works [28, 60, 43] develop advanced tokenization techniques and explore the trade-off between codebook size and codebook embedding dimension. Building on these powerful image tokenizers, recent advancements [50, 53] have
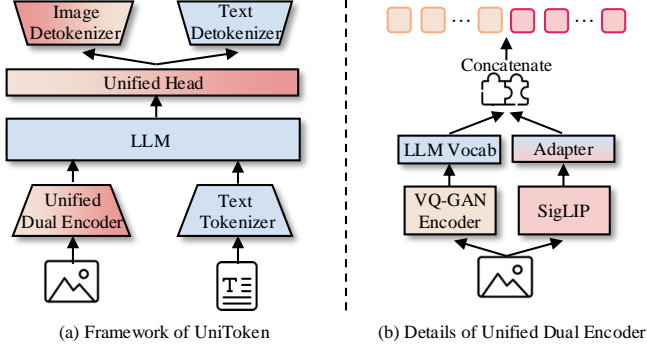
(a) Framework of UniToken   (b) Details of Unified Dual Encoder

Figure 2. Illustration of **(a) the overall framework of UniToken** and **(b) detailed designs of the unified dual encoder presented in (a)**. In (a), the "image detokenizer" and "text detokenizer" are responsible for converting predicted token IDs back into images and words, respectively. In (b), the VQ-GAN encoder processes an image and outputs discretized token IDs, which are then transformed into high-dimensional embeddings by indexing the LLM's vocabulary.

scaled up training data and transformer parameters, revealing scaling properties similar to those observed in the LLM field. However, as the tokens within the codebook are designed primarily for image reconstruction and generation tasks, they predominantly capture low-level visual details while overlooking high-level semantics, posing significant challenges for extending their application to the multimodal comprehension domain.

2.3. Unified Understanding and Generation

Unifying multimodal understanding and generation within a shared model facilitates the seamless association between high-level content reasoning with low-level image generation capabilities, allowing them to complement and enhance each other. Chameleon [41] and Lumina-mGPT [32] employ the VQ-GAN's image tokenizer as the visual encoder, and jointly train the LLM with VQA and image generation data. However, their multimodal understanding performances lag significantly behind conventional MLLMs like LLaVA-v1.5 [33] and Qwen-VL [4]. To address this, VILA-U [56] distills high-level semantics from CLIP when training its own image tokenizer. Emu3 [54] also trains a private image tokenizer based on MoVQGAN [64], which has a lower compression rate (8×8) compared with previous tokenizers. Recently, Janus [55] tackles this problem with a more straightforward design, which decouples the visual encoding of understanding and generation with the image tokenizer from SigLIP [63] and LlamaGen [50], respectively. Meanwhile, it employs two decoupled prediction heads for the two tasks. While effective, this decoupled design imposes a significant mode-switching burden when scaling to more complex tasks. In contrast, our UniToken adopts a unified visual en-

coding and prediction head, irrespective of task types.

3. UniToken

3.1. Architectural Designs

The overall framework of UniToken is illustrated in Fig. 2(a). Given a multimodal input sequence, textual inputs are tokenized using the LLM's tokenizer, while visual inputs are encoded through a dual visual encoder to produce a unified representation of continuous and discrete visual tokens. These tokens are subsequently fed into the LLM, which generates output tokens corresponding to either images or text. Finally, the output tokens are processed by their respective de-tokenizers to produce the final results. In the following section, we will elaborate on the designs of unified visual encoding and the advanced techniques, both of which are essential for developing a robust unified multimodal understanding and generation model.

Unified Visual Encoding. As illustrated in the Fig.2(b), the dual visual encoder comprises a SigLIP [63] and a VQ-Tokenizer [16] from Chameleon [41]. SigLIP ViT extracts continuous image features that are semantically rich, which are then aligned with the LLM's input space via a two-layer MLP. Meanwhile, the VQ-Tokenizer discretizes an image into a sequence of codebook IDs, which are subsequently used to retrieve their corresponding high-dimensional features from the LLM's vocabulary. Combining the continuous and discrete image features, a multimodal input sequence can be formulated as follows:

$$[\text{BOS}][\text{BOI}]\{\text{image\_d}\}[\text{SEP}]\{\text{image\_c}\}[\text{EOI}]\{\text{text}\}[\text{EOS}]$$

where $[\text{BOS}]$ and $[\text{EOS}]$ denote special tokens that mark the beginning and end of the entire sequence, respectively, while $[\text{BOI}]$ and $[\text{EOI}]$ represent special tokens that indicate the start and end of the image segment. Additionally, $[\text{SEP}]$ serves as a special token for separating the discrete ($\{\text{image\_d}\}$) and continuous image features ($\{\text{image\_c}\}$). Carrying both high-level semantics and low-level details of the input image, the unified visual encoding facilitates the LLM to selectively assimilate knowledge to deal with different tasks.

Advanced Techniques. To further enhance visual knowledge, we integrate two state-of-the-art off-the-shelf techniques to augment the unified visual encoding features. (1) To scale the continuous representation to a higher resolution, we partition the image into multiple grids and encode each grid independently as done in [33]. We choose the grid configuration of $\{2 \times 2, 1 \times \{2, 3\}, \{2, 3\} \times 1\}$ to support image inputs of varying shapes. (2) To dynamically adjust continuous representation end-to-end, we finetune the SigLIP ViT throughout our training process. Empirically, we observe that a large learning rate causes the ViT to collapse. To address this, we carefully regulate the learning

Task	Type	Dataset
Visual Und.	Caption	ShareGPT4V(200K) [8], SA1B-DenseCap(100K) [12]
	GeneralQA	LLaVA-1.6(1.5M) [33], M3IT(920K) [30], ALLaVA(1.35M) [6]
	Doc/Chart/Math	ChartQA(390K) [44], UReader(1.1M) [58], ScigraphQA(296K) [31], Geo170K(118K) [18]
	OCR	TextOCR(25K) [49], PDFA(3M) [1], SynthDoG(800K) [27], WebSRC(100K) [11]
Image Gen.	Text2Img Gen.	MidJouney(10M)

Table 1. **Distribution of training data in Stage II.** Except "MidJorney(10M)", all datasets are publicly available for academic purpose.

rate magnitude for the ViT, resulting in substantial performance improvements across a wide range of benchmarks.

3.2. Training Recipes

To effectively harmonize visual understanding and image generation capabilities, we train UniToken using a two-stage pipeline. In the following sections, we provide a detailed explanation of the training procedures for each stage.

Stage I: We utilize Chameleon [41] as our foundation model, which inherently supports image discretization and the distribution modeling of these discretized image tokens. To align the continuous image features with the base model, we freeze the LLM parameters and train only the SigLIP ViT and adapter components during Stage I. During Stage I, we utilize a dataset of 2.5 million image captions, composed of data from ShareGPT4V (49.5%), LLaVA (22.2%), and ALLaVA (28.2%).

Stage II: This stage aims to develop both visual understanding and image generation capabilities for the LLM with the aid of the unified visual encoding features. Toward this end, we train all parameters, including ViT, adapter and the LLM, during Stage II. As depicted in Tab.1, in Stage II, we utilize a dataset comprising 10 million image-to-text understanding samples and 10 million text-to-image generation samples, totaling 20 million data points. The image-to-text understanding dataset encompasses image captions, general question-answer pairs, documents, charts, mathematical problems, and OCR instances, all of which are publicly available. Additionally, the text-to-image generation data are manually synthesized by prompting the Midjourney model. The ratio between multimodal understanding and generation data is approximately 1:1. We further investigate the impact of this ratio across different data scales in Sec. 4.3.

3.3. Training Objectives

Following the prior auto-regressive prediction-based approaches, we adopt cross-entropy as the loss function:

$$\mathcal{L} = -\sum_i \log P_\theta(x_i | x_{<i}) \tag{1}$$

where θ denotes the model parameters. For the visual understanding task, we compute the loss exclusively over the

	Stage I	Stage II
Global Learning Rate	$1e^{-3}$	$1e^{-5}$
ViT Learning Rate	$1e^{-5}$	$1e^{-5}$
LR Scheduler	Cosine	Cosine
Optimizer	AdamW($\beta_1 = 0.9, \beta_2 = 0.95$)	
Batch Size	128	128
Epoch	1	1

Table 2. **Details of training hyperparameters**. "Global Learning Rate" is applied to all model parameters except for the ViT.

text tokens corresponding to the answer. For the image generation task, we calculate the loss solely over the discrete image tokens.

3.4. Inference

During inference, we employ greedy decoding to ensure deterministic outputs for the visual understanding task, while multinomial sampling decoding is utilized to enhance the diversity of generated images. Additionally, we also employ the classifier-free guidance (CFG) for image generation task following prior works [54, 55], and set the guidance scale as 5.5 and 5.0 when evaluating GenEval [21] and T2i-Compbench++ [23], respectively.

4. Experiments

In this section, we first detail the implementation specifics of UniToken, encompassing its architectural design and hyper-parameter configurations. Next, we conduct a comprehensive comparison of UniToken against a diverse range of state-of-the-art methods across both visual understanding and image generation benchmarks. Subsequently, we perform ablation studies to dissect the contributions of individual components within UniToken, followed by additional explorations to provide deeper insights. Finally, we present extensive qualitative results to visually demonstrate the capabilities and effectiveness of UniToken.

4.1. Implementation Details

For the LLM component, we adopt the architecture of Chameleon [41] and initialize its parameters using the pre-trained checkpoint from Lumina-mGPT [32]. Therefore, our UniToken inherits the discrete vision tokenizer

Method	LLM	MMMU	MMB-CN	MMB-EN	MMStar	SEED	Math-Vista	HBench	AI2D	OCRBench	MMVet
Understanding-Only											
InstructBLIP [13]	Vicuna-7B	30.6	23.9	36.0	-	53.4	25.3	45.3	-	276	26.2
QwenVL-Chat [4]	Qwen-7B	35.9	56.3	60.6	**37.5**	65.4	-	39.2	45.9	488	-
mPLUG-Owl2 [59]	LLaMA2-7B	32.7	60.7	64.5	-	57.8	22.2	-	-	255	36.5
LLaVA-v1.5 [33]	Vicuna-7B	35.3	46.4	64.3	30.3	64.3	-	**46.9**	54.8	318	30.5
ShareGPT4V [8]	Vicuna-7B	**37.2**	60.7	**68.8**	33.0	-	-	-	**58.0**	371	37.6
DeepSeek-VL [38]	DeepSeek-1B	32.2	**62.9**	64.6	-	**66.7**	-	27.6	-	409	34.8
LLaVA-v1.6(HD) [34]	Vicuna-7B	35.1	62.3	67.4	-	64.7	**34.6**	-	66.6*	**532**	**43.9**
Understanding & Generation											
Chameleon [41]	-	**34.4**	26.0	31.6	30.3	47.4	13.6	17.8	45.6	18	6.1
Lumina-mGPT [32]	Chameleon7B	25.1	16.1	33.0	28.2	51.3	21.2	35.8	44.1	37	14.0
Show-o [57]	Phi1.5-1.3B	25.1	-	-	-	-	-	-	-	-	-
Emu3† [54]	-	31.6	47.6	58.5	-	68.2	-	-	70.0*	**687**	**37.2**
VILA-U [56]	LLaMA2-7B	-	-	-	-	59.0	-	-	-	-	33.5
Janus [55]	DeepSeek-1B	30.5	52.2	69.4	-	63.7	-	-	-	-	34.3
UniToken (Ours)	Chameleon-7B	34.2	**64.1**	**70.3**	**45.2**	**70.3**	41.6	**49.6**	67.6	634	32.1

Table 3. **Comparison with state-of-the-art methods on prevalent multimodal comprehension benchmarks.** "*Understanding-Only*" refers to methods that exclusively support the multimodal understanding task, while "*Understanding & Generation*" refers to methods that support both multimodal understanding and generation tasks. "†" indicates that Emu3 trains separate models for each of the two tasks individually, and we report the results of its chat version. "*" indicates that the images of related training datasets are observed during training. "-" denotes that the results of the approaches are not reported in their paper.

Type	Method	Overall	Single Obj.	Two Obj.	GenEval Counting	Colors	Positions	Color Attri.	T2I-CompBench++ Color	Shape	Texture
Generation-Only											
Diffusion-based	DALL-E2 [47]	0.52	0.94	0.66	0.49	0.77	0.10	0.19	0.5750	0.5464	0.6374
	DALL-E3 [5]	0.67	0.96	0.87	0.47	0.83	**0.43**	0.45	**0.8110**	**0.6750**	**0.8070**
	SDv1.5 [48]	0.43	0.97	0.38	0.35	0.76	0.04	0.06	0.3730	0.3646	0.4219
	SDv2.1 [48]	0.50	0.98	0.51	0.44	0.85	0.07	0.17	0.5694	0.4495	0.4982
	SDXL [45]	0.55	0.98	0.74	0.39	0.85	0.15	0.23	0.6369	0.5408	0.5637
	SD3 [15]	**0.74**	**0.99**	**0.94**	**0.72**	**0.89**	0.33	**0.60**	-	-	-
	PixArt-α [7]	0.48	0.98	0.50	0.44	0.80	0.08	0.07	0.6886	0.5582	0.7044
AR based	LlamaGen [50]	0.32	0.71	0.34	0.21	0.58	0.07	0.04	-	-	-
Understanding & Generation											
Diffusion & AR hybrid	Show-o [57]	0.53	0.95	0.52	**0.49**	**0.82**	0.11	**0.28**	-	-	-
	SEED-X [20]	0.49	**0.97**	**0.58**	0.26	0.80	**0.19**	0.14	-	-	-
	TransFusion [65]	**0.63**	-	-	-	-	-	-	-	-	-
AR based	Chameleon [41]	0.39	-	-	-	-	-	-	-	-	-
	Emu3† [54]	0.54	**0.98**	0.71	0.34	0.81	0.17	0.21	0.6107	0.4734	0.6178
	Janus [55]	**0.61**	0.97	0.68	0.30	**0.84**	**0.46**	**0.42**	-	-	-
	UniToken (Ours)	0.58	**0.98**	**0.75**	**0.39**	0.81	0.26	0.32	**0.7115**	**0.5179**	**0.6670**

Table 4. **Comparison with state-of-the-art methods on prevalent multimodal generation benchmarks.** "*Generation-Only*" refers to methods that exclusively support image generation task. We also categorize methods based on their generation mechanisms, namely diffusion-based, autoregressive-based (AR), andhybrid approaches combining the two. "†" indicates that we use Emu3 generation version for evaluation. "-" denotes that the results of the approaches are not reported in their paper.

of Chameleon, which has a codebook of size 16,384 and downsample ratio of 16. For the continuous visual encoder, we utilize SigLIP-SO400M-Patch14-384 [63]. By further applying the aforementioned resolution scale-up technique on the SigLIP, our UniToken achieves a maximum input resolution of 768×768. The detailed hyperparameters during the whole training stages are listed in Tab.2. During the experiment, we found that using a larger ViT learning rate (e.g., $5e^{-4}$) causes the SigLIP ViT to collapse severely.

4.2. Comparison with State-of-the-Arts

Performances of Multimodal Understanding. We compare UniToken with MLLMs specialized exclusively for visual understanding tasks and MLLMs capable of both visual understanding and image generation tasks across widely recognized benchmarks, as detailed in Tab. 3. On the one hand, compared with understanding-only specialists, our UniToken achieves better performances across most bench-

#	Model	Training Data		Visual Understanding		Image Generation	
		Visual Und.	Image Gen.	General	TextRead	GenEval	T2I-Compbench++
1	Chameleon†	10M	/	45.4	28.7	-	-
2		/	10M	-	-	51.2	40.1
3		10M	10M	45.6(↑0.2)	28.4(↓0.3)	31.1(↓20.1)	26.0(↓14.1)
4	UniToken*	10M	/	53.2	50.1	-	-
5		/	10M	-	-	51.1	39.5
6		10M	10M	53.3(↑0.1)	49.5(↓0.6)	49.4(↓1.7)	39.4(↓0.1)

Table 5. **Impact of visual encoding approaches on the interference between visual understanding and image generation tasks.** "†" represents that we adopt Chameleon architecture as the discrete-only visual encoding competitor, and train it with our curated dataset for rigorous ablation. "*" indicates that input resolution scale-up technique is not employed. "/" signifies that no training data from the corresponding task types is used. "-" denotes that the performance metrics are omitted from evaluation, as no data relevant to the corresponding tasks is included in the training set. "General" and "TextRead" share the same meaning as in Tab.6.

#	Training Data		Visual Understanding		Image Generation	
	Visual Und.	Image Gen.	General	TextRead	GenEval	T2I-Compbench++
1	2M	2M	49.8	42.9	51.3	40.3
2	2M	1M	50.0(↑0.2)	43.0(↑0.1)	50.7(↓0.6)	39.7(↓0.6)
3	10M	10M	53.3	49.5	49.4	39.4
4	10M	5M	53.0(↓0.3)	49.4(↓0.1)	38.4(↓11.0)	33.8(↓5.6)

Table 6. **Impact of training data scale on the proportion of visual understanding and image generation data.** For robust evaluation, we compute the average scores across MMMU [62], MMB-CN [36], MMB-EN [36], MME [17], MMStar [9], SEED [29], MathVista [40], and HallusionBench [22], with the aggregated result denoted as "General". Similarly, we calculate the average scores across AI2D [26], OCRBench [37], and MMVet [61], with the aggregated result denoted as "TextRead". For GenEval [21] and T2I-Compbench++ [23], we report their overall scores.

marks. Notably, even when compared to LLaVA-v1.6(HD), a leading MLLM in the research field, UniToken achieves superior performance by a significant margin (e.g., +5.6 on SEEDBench, +7.0 on MathVista, and +102 on OCR-Bench). On the other hand, when compared to models capable of unified understanding and generation, UniToken also demonstrates significant performance improvements. It is noteworthy that, despite inheriting the pre-trained checkpoint from Lumina-mGPT, which lacks robust visual understanding capabilities, UniToken still outperforms leading approaches such as Emu3 (chat version) and Janus.

Performances of Image Generation. We also compare UniToken with image generation models and approaches with both visual understanding and image generation capabilities as demonstrated in Tab.4. Since diffusion-based and autoregressive methods are prominent paradigms in image generation, we further categorize approaches according to their generation paradigms. From the Tab.4, we have the following observations: (1) For image generation models, leading diffusion-based approaches, such as SD3 and DALL-E3, significantly outperforms the autoregressive-based approaches. (2) Although without enjoying the benefits of pre-trained diffusion models, our UniToken achieves competitive or even superior performances when compared with diffusion and autoregressive hybrid approaches. (3)

UniToken demonstrates performance comparable to other autoregressive-based unified models. Across specific evaluation metrics, our model achieves higher scores in categories such as "Two Obj." and "Counting", but underperforms in "Positions" and "Color Attri.". This discrepancy may stem from the absence of relevant text prompts in our text-to-image generation dataset.

4.3. Comprehensive Analysis

In this section, we delve into the fundamental design principles of developing a unified model supporting visual understanding and image generation tasks. Specifically, we first examine the interference between these two tasks and then compare the extent of such interference when employing different visual encoding techniques. Subsequently, we investigate the impact of the data proportion between these tasks on model performance and derive insights into selecting appropriate data proportions under varying training dataset scales. For efficiency, we omit the input resolution scale-up technique in all experiments conducted in this section.

Exploration of Task Interference. In this section, we train a model using three distinct datasets: visual understanding data only, image generation data only, and the combined dataset of visual understanding and image gener-

- User: "What does the image imply?"

- UniToken: "The image implies a comparison or competition between two entities, "deepseek" and "ChatGPT," which are represented by their respective logos. The use of the word "vs" suggests that the image is part of a visual representation of a match-up or contest between the two, possibly in terms of performance, features, or some other metric relevant to their services or products."

- User: " 该海报中电影的票房是多少？ "

- UniToken: " 海报上显示的电影票房为100亿元"

- User: "Please describe the details of the image as much as possible."

- UniToken: "The image is a promotional poster featuring an animated character, a young girl with black hair tied in two buns, wearing a red and gold outfit. She is looking slightly to the side with a playful expression. The background is a blend of red and white swirls, giving a sense of motion or energy. In the center, large, bold Chinese characters "81" are prominently displayed, with the number "8" in yellow and the character "1" in red. Below the number, there is a date "2025.02.10" and a time "00:00:00". At the bottom, there is a small text in red and white which appears to be a tagline or additional information, but it is not fully legible."

	Stage I	Stage II
Global Learning Rate	$1e^{-3}$	$1e^{-5}$
ViT Learning Rate	$1e^{-5}$	$1e^{-5}$
LR Scheduler	Cosine	Cosine
Optimizer	AdamW($\beta_1 = 0.9, \beta_2 = 0.95$)	
Batch Size	128	128
Epoch	1	1

Table 2. **Details of training hyperparameters**. "Global Learning Rate" is applied to all model parameters except for the ViT.

- User: " 第一阶段和第二阶段模型的全局学习率分别是多少？ "

- UniToken: "第一阶段模型的全局学习率为1e−3，第二阶段模型的全局学习率为1e−5。"

Figure 3. **The question answering results of UniToken.** Different types of questions, both in English and Chinese, are evaluated using our UniToken. Hallucinations in the responses are highlighted in red.

ation data. Meanwhile, to further investigate the effects of visual encoding approaches on task interference, we evaluate two models: (1) Chameleon [41], representing discrete-only visual encoding, and (2) our UniToken, which employs a unified discrete-continuous visual encoding strategy. As illustrated in Tab.5, for the Chameleon model, we find that joint training has little effect on the model visual understanding capability (#3 *vs.* #1), but causes severe image generation capability degradation (#3 *vs.* #2). This suggests that the discrete-only visual encoding approach struggles to effectively manage both visual understanding and image generation tasks simultaneously, with the former task tending to dominate the optimization process. On the other hand, our model demonstrates robustness in joint training compared to single-task training, both for visual understanding (#6 *vs.* #4) and image generation (#6 *vs.* #5). This high-lights that *our unified discrete-continuous visual encoding approach is less susceptible to task interference compared to the discrete-only visual encoding method*.

Exploration of Data Proportion. We ablate the effect of data proportion of visual understanding and image generation tasks under varying data scales as illustrated in Tab.6. At a smaller data scale (less than 5M), a 2:1 ratio of visual understanding to image generation data ensures the stability of both capabilities (#2 *vs.* #1). However, when scaling up the training data (more than 15M), a 2:1 ratio leads to significant degradation in image generation performance, whereas adjusting the ratio to 1:1 maintains the effectiveness of both capabilities (#4 *vs.* #3). We attribute this phenomenon to the fact that, with a 2:1 ratio, *scaling up the training data further amplifies the disparity in absolute sample sizes between visual understanding and image*

UniToken   Janus-Pro-7B

"A modern city street at night, illuminated by vibrant neon signs in Japanese kanji. In the foreground, a stoic samurai in futuristic armor with glowing katana blades stands amidst light rain, reflecting bright hues on the pavement."

UniToken   Janus-Pro-7B

"A private tropical paradise with white sandy beaches, crystal-clear turquoise waters, and an overwater villa with an infinity pool. Palm trees sway gently under a vibrant sunset."
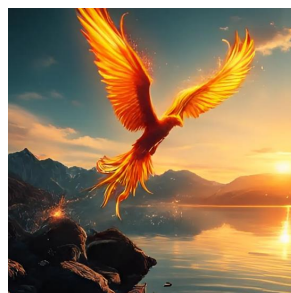
UniToken   Janus-Pro-7B

"A peaceful snow-covered valley illuminated by the soft glow of the northern lights. Frost-covered trees glisten under a starry sky, and a quaint wooden cabin with smoke rising from its chimney stands by a frozen river."

UniToken   Janus-Pro-7B

"A majestic phoenix with fiery wings soaring above a tranquil mountain lake, casting shimmer-ing reflections on the water. Sparks and embers trail behind it as the sky glows with hues of orange and gold."
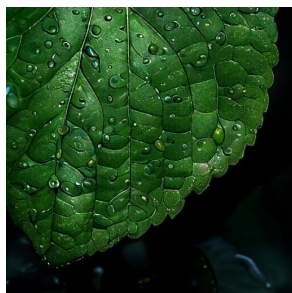
UniToken   Janus-Pro-7B
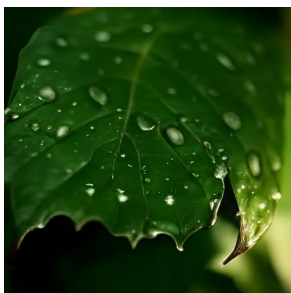
"A single dark green leaf glistening with fresh rain-drops. The veins are clearly visible, and droplets pool at the edges, refracting soft natural light."

UniToken   Janus-Pro-7B
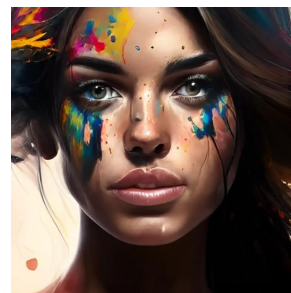
"A close-up portrait of a confident woman with vibrant paint splatters across her face. Her expressive eyes gaze directly at the viewer as dynamic brush-strokes blend surreal elements into her flowing hair."

UniToken   Janus-Pro-7B

"An old-fashioned metal lantern with a weathered bronze patina. The glass panels are slightly foggy from years of use, and a faint flickering flame glows inside, casting warm light across a dark wooden surface."

UniToken   Janus-Pro-7B

"A rustic ceramic cup with earthy brown and beige tones, featuring a textured, slightly uneven surface. Faint glaze drips add character as it sits on a stone countertop."

Figure 4. **Comparison of image generation results between UniToken and Janus-Pro-7B.**

generation tasks, ultimately resulting in the degradation of image generation performance.

4.4. Qualitative Results

Visual Understanding. For the image-to-text comprehension, we collect images of heating topics from websites and the snapshot of this paper. Afterward, we manually write questions based on the image content. As illustrated in Fig.3, our UniToken support both English and Chinese question answering, and can tackle diverse question formats to generate corresponding responses. Although some hallucinations are present in these responses, as highlighted in red in Fig.4, more advanced architectural designs and comprehensive training data aimed at enhancing visual understanding in cutting-edge MLLMs could be incorporated into our UniToken framework in the future.

Image Generation. For the text-to-image generation, we first automatically generate text prompts using the instruction of *"Please generate some text prompts for generating high-quality images"* to prompt the ChatGPT. With these text prompts, we feed them into our UniToken and the concurrent work Janus-Pro-7B [10], and demonstrate their results in Fig.4. Through comparison, it can be observed that the images generated by our UniToken exhibit finer textures and more intricate visual details than those produced by Janus-Pro-7B.

5. Discussion

Due to resource limitations, we train our UniToken model with a 7B parameter size and a mixed dataset of 20 million samples for visual understanding and image generation in this paper. In the future, we propose three potential scaling directions to further explore based on our current UniToken model: (1) **Scalability of parameter size**: Larger LLMs (e.g., the 34B parameters used in Chameleon [41]) could be adopted to investigate the potential emergent capabilities of a unified model for visual understanding and image generation. (2) **Scalability of dataset size**: Richer datasets (e.g., the nearly 200 million samples used in Janus-Pro [10]) could be utilized to push the limits of model capabilities. (3) **Scalability of task types**: A broader range of tasks (e.g., image editing, story generation) could be supported by curating corresponding training data, where our unified visual encoding, which integrates both high-level semantics and low-level details, can facilitate comprehensive joint multimodal comprehension and generation.

6. Conclusion

In this paper, we proposed UniToken, a unified visual representation that seamlessly integrates discrete and con-

tinuous tokens to bridge the gap between visual understanding and image generation tasks. Our experiments demonstrate that this approach achieves state-of-the-art performance across diverse multimodal tasks while providing valuable insights into task interference and data distribution challenges. Overall, UniToken establishes a robust foundation for future research in unified multimodal modeling.

References

[1] pdfa-eng-wds. `https://huggingface.co/datasets/pixparse/pdfa-eng-wds`.

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

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[6] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2025.

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[10] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Januspro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[11] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc:

A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021.

[12] 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.

[13] Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.

[14] 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. *arXiv preprint arXiv:2309.11499*, 2023.

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

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

[17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[18] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

[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.

[20] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[21] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

[23] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image genera-

tion. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[24] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. *arXiv preprint arXiv:2403.07304*, 2024.

[25] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

[26] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[27] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.

[28] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[29] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[30] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M³it: A large-scale dataset towards multimodal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.

[31] Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*, 2023.

[32] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

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

[34] 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.

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[36] 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 *European conference on computer vision*, pages 216–233. Springer, 2024.

[37] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[38] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024.

[40] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[41] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[42] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024.

[43] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

[44] 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. *arXiv preprint arXiv:2203.10244*, 2022.

[45] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[46] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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

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

[49] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021.

[50] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[51] 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. *arXiv preprint arXiv:2307.09288*, 2023.

[52] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[53] Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *arXiv preprint arXiv:2406.09399*, 2024.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

[59] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024.

[60] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model

beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

[61] 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. *arXiv preprint arXiv:2308.02490*, 2023.

[62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

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

[64] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022.

[65] 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.