UniModel: A Visual-Only Framework for Unified Multimodal Understanding and Generation

Chi Zhang Jiepeng Wang Youming Wang Yuanzhi Liang Xiaoyan Yang
Zuoxin Li Haibin Huang Xuelong Li
Institute of Artificial Intelligence (TeleAI), China Telecom

Abstract

We present UniModel, a unified generative model that jointly supports visual understanding and visual generation within a single pixel-to-pixel diffusion framework. Our goal is to achieve unification along three axes: the model, the tasks, and the representations.

At the representation level, we eliminate modality discrepancies by mapping both text and images into a shared visual space: textual prompts are rendered as painted text images on a clean canvas, and all inputs and outputs are treated purely as RGB pixels. This yields a fully vision-native formulation of multimodal learning.

At the task level, a broad range of vision–language problems are cast as pixel-to-pixel transformations in this visual space. For understanding tasks, the model takes an RGB image and produces a painted text image that visually encodes the semantic prediction. For generation tasks, painted text images serve as visual conditions that guide realistic and semantically aligned image synthesis. Captioning and text-to-image generation thus become different directions of the same underlying visual translation process.

At the model level, we instantiate a single Unified Diffusion Transformer trained with rectified flow in pixel space. A shared backbone jointly learns bidirectional mappings between natural images and painted text images, with lightweight task embeddings to specify the desired direction. Training is entirely pixel-based, avoiding mismatched objectives across tasks and enabling direct cross-modal supervision—for example, editing a word in the painted caption naturally steers the corresponding image region.

Experiments on text-to-image synthesis and image-to-text understanding demonstrate strong cross-modal alignment and emergent controllability such as cycle-consistent image–caption–image loops. Our initial exploration suggests that unifying model, tasks, and representations in a single visual space is a promising paradigm for general-purpose multimodal intelligence.

1. Introduction

Recent advances in multimodal AI have been driven by two parallel lines of progress: large language models (LLMs) that excel at text understanding and generation [5, 89], and diffusion-based vision models that produce high-fidelity images and videos [31, 72]. These developments have motivated a surge of "unified" architectures that aim to handle both vision and language within a single framework, including BLIP-style models [12, 42, 43], Flamingo [3], LLaVA variants [56, 57], and more recent autoregressive and diffusion-based backbones such as Kosmos-2 [68], Emu [11, 79, 98] , and VideoPoet [38].

However, despite using a shared backbone, most of these systems remain only partially unified. They typically share parameters at the model level, but still rely on heterogeneous representations and fragmented task formulations. Images and text are encoded by separate modules into different embedding spaces; understanding and generation are optimized with distinct objectives and decoders; and supervision signals are mismatched across tasks. As a result, these models often treat "image $\rightarrow$ text" and "text $\rightarrow$ image" as fundamentally different problems, which can lead to conflicting gradients, suboptimal cross-modal alignment, and limited bidirectional control.

We argue that backbone sharing alone is not enough. For truly general-purpose multimodal intelligence, a unified model should operate coherently along three axes:

(1) Representation-level unification: different modalities should be expressed in a compatible form, so that the model does not need to constantly bridge gaps between discrete symbols and continuous pixels.

(2) Task-level unification: diverse multimodal tasks should reduce to a small set of common primitives, enabling shared learning signals and consistent objectives.

(3) Model-level unification: a single parameterization should be able to support these tasks in both directions, rather than relying on separate heads or decoders for understanding vs. generation.

This perspective leads to a simple question: instead of

(a) Image-to-text (b) Text-to-image

Figure 1. Unified visual representation. (a) Given an RGB image, the model outputs its corresponding painted text image. (b) Using the painted text image as a condition, the model generates a realistic RGB image.

aligning text and images across heterogeneous spaces, can we homogenize their representation so that they can be processed as the same kind of signal?

In this work, we advocate a fully visual view of multimodal learning. We map both text and images into a shared visual space by rendering textual content as painted text images on a blank canvas. This approach builds on a growing body of work that treats text as a visual signal for tasks like language modeling [73], long-context compression [62, 100], and generative conditioning [44, 117]. Once text is represented as an RGB image, captions, questions, and answers become visual objects, just like natural images. This yields representation-level unification: all inputs and outputs are RGB images, and the model operates entirely in a visual domain. Under this representation, a wide range of vision–language tasks collapse into the same primitive: pixel-to-pixel translation. Visual understanding tasks, such as image captioning or visual question answering, can be cast as RGB-image $\rightarrow$ painted-text-image transformations. Visual generation tasks, such as text-to-image synthesis, become painted-text-image $\rightarrow$ RGB-image transformations. From this viewpoint, understanding and generation are no longer separate pipelines, but opposite directions of the same visual mapping, achieving task-level unification.

Building on this idea, we introduce UniModel, a unified diffusion framework that treats multimodal understanding and generation as a single pixel-to-pixel problem. At the model level, UniModel is instantiated as a bidirectional diffusion transformer [67] trained with a rectified-flow [59] objective on VAE latents [72]. A shared backbone jointly learns all mappings between natural and painted images, while lightweight task embeddings indicate the desired direction (e.g., understanding vs. generation). Training is performed entirely in pixel space, with consistent diffusion objectives for both directions, achieving model-level unification without task-specific decoders or losses.

This three-fold unification has several practical benefits. First, it provides a clean and symmetric formulation of image$\leftrightarrow$text relations, which encourages stronger cross-modal alignment and cycle consistency. Second, it enables natural forms of controllability: editing a word in the painted caption becomes a local visual edit that can steer the corresponding region of the generated image. Third, it simplifies the overall system design, avoiding hand-crafted interfaces between separate language and vision modules.

In summary, this work makes the following contributions:

- We propose a representation-level unification strategy that maps both text and images into a shared visual space via painted text images, enabling a fully vision-native formulation of multimodal learning.
- We cast a broad range of vision–language tasks as pixel-to-pixel translations between natural and painted images, achieving task-level unification of visual understanding and generation.
- We introduce UniModel, a single bidirectional diffusion transformer trained with rectified-flow objectives on VAE latents, which realizes model-level unification with shared parameters and lightweight task embeddings.
- We demonstrate on text-to-image synthesis and image-to-text understanding that UniModel attains competitive performance, strong cross-modal alignment, and emergent controllability such as cycle-consistent image–caption–image reconstruction.

2. Related Work

Unified models The pursuit of foundation models that holistically unify visual and textual understanding and generation has catalyzed the exploration of several architectural paradigms. A distinct approach leverages unified discrete diffusion frameworks, which model multimodal data in a non-autoregressive manner as seen in [49, 74, 80, 96, 116]. However, the predominant approach centers on extending Multimodal Large Language Models (MLLMs) with autoregressive (AR) generation capabilities, differentiated by their visual encoding strategies. One strategy involves Pixel Encoding, where images are represented by discrete visual tokens, forming a visual vocabulary for the AR model. This is exemplified by works such as [10, 11, 39, 41, 53, 58, 82, 85, 92, 98, 104, 107, 115]. A more prevalent strategy is Semantic Encoding, which utilizes semantically rich tokens derived from powerful vision encoders like CLIP. This paradigm is foundational to a vast array of models including [8, 15, 17, 23, 27, 37, 46, 48, 51, 52, 63, 76, 78, 79, 87, 88, 93, 94, 103, 108, 114, 124, 128]. Further refinement is achieved through Learnable Query Encoding, where a set of learnable vectors distill visual features into fixed-length representations, a technique central to models like [1, 2, 6, 20–22, 66, 83, 106, 113, 122]. Concurrently, Hybrid Encoding strategies combine semantic and pixel-level features through either pseudo-decoupled [7, 16, 97, 101, 109, 131] or jointly trained [9, 33, 35, 71, 111, 112, 130] visual tokenizers. Bridging these paradigms, a third family of MLLM

(AR+Diffusion) models integrates diffusion mechanisms into the AR framework for high-fidelity image decoding, as demonstrated by [13, 50, 64, 75, 110, 123, 126] with hybrid encoding. The ultimate ambition of this field is manifested in Any-to-Any models, which aim to process and generate arbitrary sequences of multimodal inputs, a frontier being explored by [26, 36, 40, 45, 61, 99, 105, 118, 121].

Unified representation via visual tokens An emergent paradigm in multimodal understanding seeks to bypass discrete tokenization by processing text as a visual modality. This principle was established by models like [73], which render text into images to enhance cross-lingual transfer. This unification is advanced by models like [90], using a single encoder for images and rendered text, and [34], which fuses visual and text prompts. In document AI, models like [84] process text and layout as a unified visual input, while frameworks like [129] train a single vision encoder for both image and text recognition. This focus on fine-grained text extends to using explicit glyph-level features as visual conditions, as seen in [24, 117]. A novel application is long-context compression, where models like [62, 91, 100] compress text into a compact image representation. Collectively, these methods underscore a significant shift towards treating language as a visual signal at the input stage.

Unified generation via visual representations The application of diffusion models to discrete text has primarily spurred two research directions: developing diffusion processes directly in the discrete token space [4, 30, 70, 127], on word embeddings for sequence-to-sequence tasks [25], and, more prominently, operating in a continuous proxy space by applying the diffusion-denoising process on word embeddings before projecting back to a vocabulary [14, 19, 25, 28, 32, 44, 47, 54, 55, 60, 77, 119, 120, 125]. Distinct from these token-based strategies, an emergent paradigm reframes text generation as a task within the visual domain, thereby circumventing the constraints of traditional vocabularies. This principle was pioneered by methods like [44], which leverages diffusion models to synthesize a complete glyph image of the target text for subsequent decoding into a symbolic sequence, and [81], which performs direct auto-regressive generation in pixel space. The concept has been extensively applied to not only font generation [18, 29, 86], but also the creation of consistent, high-fidelity text effects via shape-adaptive diffusion models like [65], where the creation of individual characters is treated as a conditional image synthesis problem solved via a diffusion process. A higher level of integration is demonstrated by large-scale multimodal models such as [11, 102], which achieve native, high-quality rendering of complex text as an integral component of the generated image within a unified framework. Furthermore, this philosophy has been operationalized through external control mechanisms; techniques based on ControlNet [69, 95], utilize visual layouts as guidance to precisely "paint" text within the latent space of an image. Our work is situated at the confluence of these advancements, employing a unified Diffusion Transformer (DiT) architecture that operates directly on pixel representations to seamlessly generate both text (as images) and photorealistic images within a single, parallel model.

3. Method

In this section, we introduce UniModel, a unified pixel-to-pixel diffusion framework that solves both visual understanding (image $\rightarrow$ text) and visual generation (text $\rightarrow$ image) within a single architecture. This section is organized as follows: Sec. 3.1 describes the unified visual representation. Sec. 3.2 introduces the bidirectional diffusion transformer. Sec. 3.3 presents the unified rectified-flow training objective. Figure 2 illustrates the overall pipeline.

3.1. Unified Visual Representation

To unify image generation and understanding within a single framework, we express all inputs and outputs—natural images as well as textual prompts—directly in pixel space. This introduces a fundamental representational shift: semantic content, including text, is modeled visually rather than symbolically. Instead of mapping text to discrete token embeddings, we render it as RGB images, enabling learning and inference to occur entirely within a single visual modality.

Text as Pixels. We convert textual sequences into RGB images by rendering each prompt, caption, question, or answer onto a blank $512 \times 512 \times 3$ canvas using a fixed font and line spacing, producing a *painted text image*. This 2D rendering preserves the underlying semantics while encoding text in a spatially structured form that is naturally compatible with both convolutional and transformer-based vision modules. For sequences that exceed the available canvas area, we truncate the text to maintain consistent input dimensionality across samples.

A Single Visual Modality. Once rendered, painted text images and natural RGB images share the same dimensionality, the same pixel-level representation, and the same preprocessing pipeline. Both types of images are encoded through the same VAE to produce latent variables that lie in a unified latent space. As a result, UniModel can treat multimodal tasks as transformations between two visual domains:

$$\text{RGB image} \longleftrightarrow \text{painted text image}.$$

(a) Image-to-text (b) Text-to-image

Figure 2. Unified bidirectional generation. (a) Image-to-text: the model encodes an input image into conditional states and generates its text-painted counterpart. (b) Text-to-Image: provided a painted text image, the model synthesizes a realistic RGB image using the same unified architecture.

This formulation offers several advantages. (1) **Unified representation**: both text and images are treated uniformly as visual signals, eliminating modality gaps between discrete symbols and continuous pixels. (2) **Unified objective**: understanding and generation share the same rectified-flow diffusion loss, removing the need for cross-entropy decoders, autoregressive language heads, or specialized text objectives. (3) **Unified architecture**: a single diffusion transformer handles all multimodal tasks bidirectionally, enabling strong cross-modal consistency, natural cycle inference, and seamless controllability (e.g., editing a word in the painted caption directly influences corresponding image regions).

3.2. Unified Diffusion Transformer

Our model builds on the MMDiT architecture used in Qwen-Image and consists of two components: (i) a visual encoder module that processes both RGB images and painted text images as model conditions, and (ii) a diffusion-based generative model that produces either RGB images or rendered text, depending on the task.

All inputs—natural or text-rendered—are encoded by a shared visual encoder into semantic hidden states, which condition the diffusion transformer. A pretrained VAE maps images to a latent space where the rectified-flow diffusion process operates. For understanding tasks, the model takes an RGB image and generates a painted text image; for generation tasks, the input is a painted text image and the output is RGB. This fully visual pipeline removes the need for modality-specific encoders or cross-modal fusion, enabling true architectural unification.

To train the model bidirectionally, we randomly swap input–output pairs (RGB $\rightarrow$ text or text $\rightarrow$ RGB) with equal probability, encouraging symmetric learning of rendering and reconstruction. To further clarify the desired direction, we introduce two learnable task embeddings—one for understanding and one for generation—which are concatenated with conditioning features before diffusion. These embeddings provide explicit task awareness, improving controllability and convergence.

3.3. Training

To train the model, we adopt *rectified flow*, which improves stability by predicting the instantaneous velocity along the optimal transport path from data to prior. Let $\mathbf{z}_0$ be the VAE latent of the target image (RGB or painted text), and let $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote a Gaussian prior. Rectified flow defines a straight-line interpolation:

$$\mathbf{z}(t) = (1-t)\mathbf{z}_0 + t\mathbf{z}_1, \quad t \in [0,1],$$

with ground-truth velocity field

$$\mathbf{v}^*(\mathbf{z}(t), t) = \mathbf{z}_1 - \mathbf{z}_0.$$

Given conditioning features $\mathbf{c}$ from the visual encoder, a task embedding $\mathbf{e}_{\text{task}}$, and time step $t \sim \mathcal{U}(0,1)$, the diffusion transformer f_θ is trained to match this velocity using the objective

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\mathbf{z}_0,\,\mathbf{z}_1,\,t}\left[\left\|\mathbf{v}^*(\mathbf{z}(t),t) - f_\theta\big(\mathbf{z}(t),\,t,\,\mathbf{c},\,\mathbf{e}_{\text{task}}\big)\right\|_2^2\right].$$

Under bidirectional training, paired RGB and painted-text latents are randomly assigned as source or target (RGB $\rightarrow$ text or text $\rightarrow$ RGB), allowing a single objective to supervise both generation and understanding.

4. Experiments

4.1. Implementation details

We build our Unified Diffusion Transformer upon the MMDiT (Multi-Modal Diffusion Transformer) architecture, inheriting its block design, attention mechanisms, and adaptive layer normalization (AdaLN) for conditional generation.

4.2. Text to image generation

For text-to-image generation, our model takes a painted text image as the conditioning input instead of raw text tokens. This design removes the need for separate text encoders or cross-modal fusion modules, allowing the entire generation pipeline to operate purely within the visual latent space. However, because existing unified or diffusion-based models rely on learned textual embeddings—and none of them accept visualized text as input—direct quantitative comparison under the same input paradigm is not feasible.

To assess the effectiveness of our approach, we provide qualitative results in Fig. 3. As shown, our model generates images that are not only visually compelling but also semantically faithful to the content encoded in the painted text prompts. The generated samples accurately capture described objects, attributes, and scene structures, demonstrating that visually represented text provides a strong and interpretable conditioning signal. Moreover, the model is robust to variations in layout, font style, and formatting in the painted text, indicating that it learns a consistent visual-language grounding within a unified representation space.

These results collectively show that visualizing text as images is a viable and effective strategy for conditioning image generation, enabling seamless integration of understanding and generation capabilities within a single diffusion transformer

4.3. Image Understanding

For image understanding tasks, our model takes an RGB image as input and predicts a painted text image that visually describes semantic outputs. This formulation eliminates the need for text decoders or multimodal fusion modules; instead, all reasoning and prediction occur entirely within the visual latent domain.

To demonstrate the effectiveness of this unified representation, we present qualitative results in Fig. 4. As shown, the model produces text outputs that are both semantically accurate and spatially well-structured. The painted predictions not only reflect correct high-level semantics—such as object descriptions, attributes, and relationships—but also exhibit coherent layout and formatting. In particular, the model autonomously organizes text spatially (e.g., line breaking and alignment), suggesting that the visual diffusion transformer implicitly learns both language semantics and text layout conventions.

These results indicate that our unified visual framework supports robust image understanding, validating the versatility of directly representing textual information in visual form.

4.4. Analysis

Thanks to its unified architecture and bidirectional training objective, UniModel naturally supports **cycle-consistent inference**: RGB $\rightarrow$ painted text $\rightarrow$ reconstructed RGB.

Given an input RGB image $\mathbf{x}_{\text{rgb}}$, the model first predicts a *painted text image*, which visually describes high-level semantics such as object categories, attributes, and spatial relations. Although this intermediate representation may contain rendering imperfections (e.g., distorted glyphs or incomplete words), it generally remains semantically faithful. Conditioning the model on $\mathbf{x}_{\text{text}}$ with task=generation then produces a reconstructed RGB image $\hat{\mathbf{x}}_{\text{rgb}}$ that preserves the original semantics while often exhibiting clean compositions and consistent visual structure. The generative process is resilient to imperfections in the painted-text input, indicating strong semantic grounding and cross-modal consistency within the shared visual space. Fig. 5 shows a gallery of cycle inference results.

These observations suggest that UniModel learns a unified visual representation in which visual and textual representations are unified in visual space—despite the absence of any explicit cycle-consistency loss during training.

4.5. Limitations

While UniModel achieves strong unification across vision understanding and generation, several limitations remain: **(1) Training efficiency.** Jointly optimizing for bidirectional tasks (RGB $\leftrightarrow$ text rendering) increases optimization complexity, leading to longer convergence times than single-direction baselines under identical hardware/resources. This stems from the need to balance gradient signals across heterogeneous objectives and stabilize the shared representation space. **(2) Text rendering fidelity.** The rendered text images—though semantically meaningful—can exhibit visual artifacts, such as font inconsisten-

Figure 3. Gallery of text-to-image generation results.

Figure 4. Gallery of image-to-textimage generation results.

Figure 5. Cycle inference results. Starting from an input image, the model generates a painted text description and then reconstructs an image from the painted text. The reconstructed image preserves the key semantics and visual attributes of the original, demonstrating strong bidirectional consistency in our unified visual representation.

cies, glyph distortions (Fig. 6). While the generative head is robust to mild imperfections (Fig. 5), severe artifacts may degrade downstream reconstruction quality or mislead fine-grained understanding.

We view these challenges as promising directions for future work: e.g., incorporating perceptual text priors, adaptive curriculum learning, or lightweight task routing to improve efficiency and fidelity.

Figure 6. Failure case. Although the description is semantically reasonable, the painted text includes incorrect characters and misspelled words, highlighting remaining challenges in accurate word-level generation.

5. Conclusion and Future Work

We introduced UniModel, a Unified Diffusion Transformer that integrates visual understanding and generation within a single, modality-agnostic framework. By representing both natural images and textual content as pixel-based inputs, UniModel removes the need for language tokenization, modality-specific encoders, and cross-modal fusion modules. Combined with bidirectional training, task-aware embeddings, and a rectified-flow objective, this unified design enables a single model to perform image–text and text–image transformations with strong cross-modal alignment and emergent properties such as cycle-consistent inference.

Despite remaining challenges—such as improving text rendering fidelity and further scaling training efficiency—UniModel demonstrates the potential of treating perception and generation as two directions of the same pixel-to-pixel mapping. We believe this work provides a step toward general-purpose visual foundation models that learn, reason, and synthesize entirely within a unified visual space, and we hope it motivates future research into fully end-to-end multimodal systems built on shared pixel-level representations.

We believe UniModel represents a step toward general-purpose visual foundation models that learn, reason, and synthesize entirely within a unified visual space, and we hope this work inspires further research into end-to-end multimodal systems grounded in shared pixel-level representations.

References

[1] Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, Guang-Ming Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua

Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-omni: A unified multimodal model for perception and generation, 2025. 2

[2] Inclusion AI, Biao Gong, Cheng Zou, Dandan Zheng, Hu Yu, Jingdong Chen, Jianxin Sun, Junbo Zhao, Jun Zhou, Kaixiang Ji, Lixiang Ru, Libin Wang, Qingpei Guo, Rui Liu, Weilong Chai, Xinyu Xiao, and Ziyuan Huang. Minglite-uni: Advancements in unified architecture for natural multimodal interaction, 2025. 2

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 1

[4] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2023. 3

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[6] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 2

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. 2

[8] Yanzhe Chen, Huasong Zhong, Yan Li, and Zhenheng Yang. Unicode^2: Cascaded large-scale codebooks for unified multimodal understanding and generation, 2025. 2

[9] Zisheng Chen, Chunwei Wang, Xiuwei Chen, Hongbin Xu, Runhui Huang, Jun Zhou, Jianhua Han, Hang Xu, and Xiaodan Liang. Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation, 2025. 2

[10] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation, 2024. 2

[11] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners, 2025. 1, 2, 3

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[13] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. 3

[14] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data, 2022. 3

[15] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Dreamllm: Synergistic multimodal comprehension and creation, 2024. 2

[16] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, Michael Rubinstein, Michalis Raptis, Deqing Sun, and Radu Soricut. Unified autoregressive visual generation and understanding with continuous tokens, 2025. 2

[17] Rongyao Fang, Chengqi Duan, Kun Wang, Hao Li, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Hongsheng Li, and Xihui Liu. Puma: Empowering unified mllm with multi-granular visual generation, 2024. 2

[18] Bin Fu, Fanghua Yu, Anran Liu, Zixuan Wang, Jie Wen, Junjun He, and Yu Qiao. Generate like experts: Multi-stage font generation by incorporating font transfer process into diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6892–6901, 2024. 3

[19] Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Empowering diffusion models on the embedding space for text generation, 2024. 3

[20] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model, 2023. 2

[21] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer, 2023.

[22] 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, 2025. 2

[23] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu,

Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again, 2025. 2

[24] Syeda Anshrah Gillani, Mirza Samad Ahmed Baig, Osama Ahmed Khan, Shahid Munir Shah, Umema Mujeeb, and Maheen Ali. Textpixs: Glyph-conditioned diffusion with character-aware attention and ocr-guided supervision, 2025. 3

[25] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models, 2023. 3

[26] Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, Ming Yang, and Jun Zhou. M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance, 2025. 3

[27] Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations, 2025. 2

[28] Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control, 2023. 3

[29] Haibin He, Xinyuan Chen, Chaoyue Wang, Juhua Liu, Bo Du, Dacheng Tao, and Yu Qiao. Diff-font: Diffusion model for robust one-shot font generation, 2023. 3

[30] Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models, 2022. 3

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1

[32] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions, 2021. 3

[33] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, and Hang Xu. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement, 2025. 2

[34] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy, 2024. 3

[35] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding, 2025. 2

[36] Yang Jin, Zhicheng Sun, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, Kun Gai, and Yadong Mu. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization, 2024. 3

[37] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified language-vision pretraining in llm with dynamic discrete visual tokenization, 2024. 2

[38] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024. 1

[39] Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads, 2025. 2

[40] Jinxiang Lai, Jie Zhang, Jun Liu, Jian Li, Xiaocheng Lu, and Song Guo. Spider: Any-to-many multimodal llm, 2025. 3

[41] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding, 2024. 2

[42] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1

[44] Junyi Li, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Glyphdiffusion: Text generation as image generation, 2023. 2, 3

[45] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows, 2025. 3

[46] Teng Li, Quanfeng Lu, Lirui Zhao, Hao Li, Xizhou Zhu, Yu Qiao, Jun Zhang, and Wenqi Shao. Unifork: Exploring modality alignment for unified multimodal understanding and generation, 2025. 2

[47] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation, 2022. 3

[48] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models, 2024. 2

[49] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding, 2025. 2

[50] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation, 2025. 3

[51] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang,

Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 2

[52] Han Lin, Jaemin Cho, Amir Zadeh, Chuan Li, and Mohit Bansal. Bifrost-1: Bridging multimodal llms and diffusion models with patch-level clip latents, 2025. 2

[53] Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation, 2025. 2

[54] Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise, 2023. 3

[55] Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. Composable text controls in latent space with odes, 2023. 3

[56] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

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

[58] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2025. 2

[59] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 2

[60] Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q. Weinberger. Latent diffusion for language generation, 2023. 3

[61] 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, 2023. 3

[62] Yujie Lu, Xiujun Li, Tsu-Jui Fu, Miguel Eckstein, and William Yang Wang. From text to pixel: Advancing long-context understanding in mllms, 2024. 2, 3

[63] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding, 2025. 2

[64] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2025. 3

[65] Xinzhi Mu, Li Chen, Bohan Chen, Shuyang Gu, Jianmin Bao, Dong Chen, Ji Li, and Yuhui Yuan. Fontstudio: Shape-adaptive diffusion model for coherent and consistent font effect generation, 2024. 3

[66] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. 2

[67] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 2

[68] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. 1

[69] KhayTze Peong, Seiichi Uchida, and Daichi Haraguchi. Typographic text generation with off-the-shelf diffusion model, 2024. 3

[70] Lihua Qian, Mingxuan Wang, Yang Liu, and Hao Zhou. Diffusion glancing transformer for parallel sequence to sequence learning, 2023. 3

[71] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation, 2025. 2

[72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[73] Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels, 2023. 2, 3

[74] Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model, 2025. 2

[75] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation, 2025. 3

[76] Wei Song, Yuran Wang, Zijia Song, Yadong Li, Haoze Sun, Weipeng Chen, Zenan Zhou, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies, 2025. 2

[77] Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. Self-conditioned embedding diffusion for text generation, 2022. 3

[78] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners, 2024. 2

[79] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality, 2024. 1, 2

[80] Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion, 2025. 2

[81] Yintao Tai, Xiyang Liao, Alessandro Suglia, and Antonio Vergari. Pixar: Auto-regressive language modeling in pixel space, 2024. 3

[82] Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning, 2025. 2

[83] Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing, 2025. 2

[84] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing, 2023. 3

[85] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. 2

[86] Vikas Thamizharasan, Difan Liu, Shantanu Agarwal, Matthew Fisher, Michael Gharbi, Oliver Wang, Alec Jacobson, and Evangelos Kalogerakis. Vecfusion: Vector font generation with diffusion, 2024. 3

[87] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer, 2024. 2

[88] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning, 2024. 2

[89] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1

[90] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only, 2023. 3

[91] Alex Jinpeng Wang, Linjie Li, Yiqi Lin, Min Li, Lijuan Wang, and Mike Zheng Shou. Leveraging visual tokens for extended text contexts in multi-modal learning. *NeurIPS*, 2024. 3

[92] Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, Mingze Zhou, Wang Lin, Kaihang Pan, Saining Zhang, Liyu Jia, Wentao Hu, Wei Zhao, and Hanwang Zhang. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning, 2025. 2

[93] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance, 2024. 2

[94] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report, 2025. 2

[95] Haofan Wang, Yujia Xu, Yimeng Li, Junchen Li, Chaowei Zhang, Jing Wang, Kejia Yang, and Zhibo Chen. Reptext: Rendering visual text via replicating, 2025. 3

[96] Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. Fudoki: Discrete flow-based unified understanding and generation via kinetic-optimal velocities, 2025. 2

[97] Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, Hongyang Wei, Eric Li, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork unipic: Unified autoregressive modeling for visual understanding and generation, 2025. 2

[98] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. 1, 2

[99] Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, Haoran Que, Zhaoxiang Zhang, Yuanxing Zhang, Ge Zhang, Ke Xu, Jie Fu, and Wenhao Huang. Mio: A foundation model on multimodal tokens, 2025. 3

[100] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression, 2025. 2, 3

[101] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024. 2

[102] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 3

[103] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025. 2

[104] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators, 2025. 2

[105] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2024. 3

[106] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation, 2025. 2

[107] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation, 2025. 2

[108] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2025. 2

[109] Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language models with rgpo, 2025. 2

[110] 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, 2025. 3

[111] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models, 2025. 2

[112] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding, 2025. 2

[113] Junzhe Xu, Yuyang Yin, and Xi Chen. Tbac-uniimage: Unified understanding and generation by ladder-side diffusion tuning, 2025. 2

[114] Zhiyang Xu, Jiuhai Chen, Zhaojiang Lin, Xichen Pan, Lifu Huang, Tianyi Zhou, Madian Khabsa, Qifan Wang, Di Jin, Michihiro Yasunaga, Lili Yu, Xi Victoria Lin, and Shaoliang Nie. Pisces: An auto-regressive foundation model for image understanding and generation, 2025. 2

[115] Jian Yang, Dacheng Yin, Yizhou Zhou, Fengyun Rao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Mmar: Towards lossless multi-modal auto-regressive probabilistic modeling, 2025. 2

[116] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models, 2025. 2

[117] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation, 2023. 2, 3

[118] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, and Hongxu Yin. X-vila: Cross-modality alignment for large language model, 2024. 3

[119] Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. Dinoiser: Diffused conditional sequence learning by manipulating noises, 2024. 3

[120] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Seqdiffuseq: Text diffusion with encoder-decoder transformers, 2023. 3

[121] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. Anygpt: Unified multimodal llm with discrete sequence modeling, 2025. 3

[122] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. Nexus-gen: Unified image understanding, generation, and editing via prefilled autoregression in shared embedding space, 2025. 2

[123] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression, 2024. 3

[124] Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies autoregressive multimodal understanding and generation, 2025. 2

[125] Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation, 2024. 3

[126] 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, 2024. 3

[127] Kun Zhou, Yifan Li, Wayne Xin Zhao, and Ji-Rong Wen. Diffusion-nat: Self-prompting discrete diffusion for non-autoregressive text generation, 2023. 3

[128] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation, 2023. 2

[129] Yi Zhu, Yanpeng Zhou, Chunwei Wang, Yang Cao, Jianhua Han, Lu Hou, and Hang Xu. Unit: Unifying image and text recognition in one vision encoder, 2024. 3

[130] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model, 2025. 2

[131] Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models, 2025. 2