

A Related Work

A.1 Unified multimodal LLMs

Autoregressive Paradigms: End-to-End and Two-Stage Modeling. Autoregressive (AR) modeling remains a core strategy for unified multimodal understanding and generation, but recent advances have led to two distinct AR-based paradigms.

The first is the *end-to-end AR paradigm*, in which all modalities—including images, text, video, and even audio—are tokenized into a unified discrete space and directly modeled within a single AR sequence framework. Representative works such as Unified-IO [101, 102], Chameleon [54], AnyGPT [103], and Emu3 [18] follow this approach: a transformer autoregressively predicts the next token across modalities, with image tokens directly decoded back to pixels via learned decoders such as VQGAN. DDT-Llama [104] further improves tokenization by introducing recursive diffusion timestep tokens, enabling better alignment with language modeling and image reconstruction. This approach enables strong performance in both understanding and generation, and supports flexible modality conversion (e.g., AnyGPT covers speech and music). Building on this foundation, models like Janus [20] and Janus-Pro [26] decouple visual encoding for understanding and generation to address the granularity mismatch, while VILA-U [90], LWM [69], and LaVIT [55] focus on efficient tokenization, unified visual-text alignment, and scaling to long-context and video scenarios. Illume [21] and Illume+ [52] further enhance data efficiency and token alignment, with Illume+ introducing dual visual tokenization and a diffusion-based decoder for higher-fidelity image synthesis and editing.

By contrast, the *two-stage AR+diffusion paradigm* separates sequence modeling and image synthesis: AR models first generate image tokens, which are then used as conditions for downstream diffusion decoders to boost image quality and diversity. Representative works include DreamLLM [105], which enables free-form interleaved multimodal generation; MiniGPT-5 [106], which improves image-text coherence with a two-stage pipeline; NExT-GPT [107], which supports any-to-any modality conversion by connecting AR sequence modeling with modular diffusion decoders; MetaMorph [88], which efficiently adapts LLMs for unified text and visual token generation; SEED-LLaMA [17], which aligns image token semantics with text for scalable multimodal autoregression; and SEED-X [73], which further enables arbitrary-size and multi-granularity image generation. Recently, BLIP3-o [108] advanced this paradigm by generating CLIP-based image features using a diffusion transformer and adopting sequential pretraining to better balance understanding and generation. Collectively, these models demonstrate the flexibility and high image fidelity achievable with the two-stage approach, highlighting a distinct trade-off with end-to-end AR models in reasoning and generation quality.

Hybrid Paradigm: Integrating AR and Diffusion within a Unified Framework. To bridge the gap between the reasoning strengths of AR models and the generative power of diffusion models, hybrid paradigms have emerged that combine both mechanisms in a unified architecture. For example, JanusFlow [109] employs a continuous reactified flow for image generation, Show-o [56] adopts a discrete MaskGIT-style diffusion, while Transfusion [19] utilizes a continuous U-Net-based DDPM. Despite their differences in diffusion implementation, these hybrid models all enable more flexible and controllable vision-language generation, further blurring the boundaries between AR and diffusion approaches.

Diffusion Paradigm: Fully Diffusion-Based Multimodal Generation. In parallel, fully diffusion-based approaches have also been proposed for unified multimodal modeling. UniDisc [48] and D-Dit [46] formulate both text and image generation as a discrete diffusion process, starting from masked sequences and enabling joint inpainting, editing, and controllable multimodal generation. By leveraging the iterative denoising process, diffusion models typically achieve superior generation fidelity and support fine-grained, high-quality editing. Moreover, unlike autoregressive models that generate tokens sequentially, diffusion-based approaches can produce multiple tokens in parallel during inference, improving efficiency and enabling more globally consistent outputs. While these models offer enhanced controllability and flexible inference, they may still face challenges in complex instruction following and sequential reasoning. Nevertheless, fully diffusion-based paradigms represent a promising direction for scenarios requiring fine-grained editing, state-of-the-art generation quality, and efficient parallel decoding across modalities.

Comparisons with Bagel [25]. Bagel [25] is a very strong recent advance in unified multimodal understanding and generation. While both FUDOKI and Bagel aim for unified multimodal mod-

eling, they are based on fundamentally different generative paradigms and architectural choices. Specifically, Bagel employs a large Mixture-of-Transformer-Experts (MoT) architecture and follows the autoregressive (AR) modeling paradigm, enabling it to efficiently scale with massive, carefully structured interleaved multimodal data. In contrast, FUDOKI is the first general-purpose unified multimodal model built entirely on discrete flow matching, which allows for bidirectional information integration and iterative self-correction during generation. In terms of empirical performance, Bagel demonstrates strong results on both multimodal generation and understanding, including advanced tasks such as free-form image manipulation. We acknowledge that FUDOKI currently lags behind Bagel, which can be attributed mainly to Bagel’s novel data scaling strategies and substantially larger model size (14B parameters for Bagel vs. 1.5B for FUDOKI). We will explore integrating similar scaling approaches in future work.

A.2 Flow Matching

Flow matching offers a fundamentally different approach to generative modeling compared to diffusion models. While diffusion models rely on repeatedly injecting random noise into data and then iteratively denoising it, flow matching instead learns a smooth, continuous transformation, formulated through ordinary differential equations (ODEs), that maps a simple distribution (such as Gaussian noise) directly to real data. This approach eliminates the need for repeated noise addition and removal.

Pioneering this direction, Lipman et al. [42] introduced Continuous Normalizing Flows (CNFs) and the flow matching framework, which trains neural networks by regressing vector fields along flexible probability paths. This work laid the foundation for subsequent advances in CNF-based generative modeling. Building on this, Liu et al. [41] proposed Rectified Flow, which learns neural ODEs along straight-line paths between distributions, enabling more efficient and scalable training for tasks such as image generation and domain adaptation. More recently, Albergo and Vanden-Eijnden [110] presented InterFlow, which simplifies training by directly inferring the velocity field from the probability flow of an interpolant density, thus avoiding costly ODE backpropagation and supporting efficient likelihood estimation and high-resolution generation.

A key advantage of flow matching is its **sampling efficiency**: by allowing deterministic sampling in just a few ODE steps, it achieves competitive FID scores with orders of magnitude fewer steps compared to diffusion-based samplers. This remarkable efficiency has quickly made flow matching a dominant approach in state-of-the-art image and video generation models.

Recent studies have also extended flow matching to discrete data domains. Campbell et al. [39] introduced Discrete Flow Models (DFMs), which generalize flow matching to discrete spaces using continuous-time Markov chains, improving multimodal modeling of both continuous and discrete data over discrete diffusion models. Similarly, Gat et al. [37] proposed Discrete Flow Matching, a framework that supports general probability paths and scalable non-autoregressive generation, significantly narrowing the performance gap between discrete flow and autoregressive models on coding benchmarks.

Thanks to these advances, flow matching methods have demonstrated strong performance across a wide range of domains, including image synthesis [14, 15], video generation [111–114], speech and audio generation [115–117], protein design [118–120], and robot control [121]. These successes underscore the broad applicability and effectiveness of flow matching frameworks.

A.3 Discrete Diffusion Models

Diffusion models have achieved remarkable success in continuous domains such as images and audio [57, 122, 123]. However, their adaptation to natural language poses unique challenges due to the discrete nature of text. Early attempts to overcome this primarily injected Gaussian noise into token embedding spaces, followed by denoising to reconstruct discrete sequences [124, 125]. Representative models in this line include Diffusion-LM [124], DiffuSeq [125], and Plaid [126]. While these approaches show promise for controllable generation and sequence-to-sequence tasks, the need to map between discrete and continuous representations complicates training and inference.

Recent research has shifted to discrete noise-based diffusion models to address these limitations, where noise injection and denoising are directly defined in the symbol space. The most influential

early works in this direction are Argmax Flows [127] and D3PM [33]. D3PM, in particular, provides a systematic framework for discrete diffusion, formalizing both absorbing (mask-based) and uniform (categorical) noise processes for sequence corruption. These foundational studies enable the progressive corruption of discrete sequences through distinct forward processes: in the absorbing (mask-based) process, tokens in the original sequence are gradually replaced with a special absorbing token (e.g., <MASK>); in the uniform (categorical) process, tokens are progressively replaced with randomly sampled tokens from the vocabulary. The diffusion model is then trained to reverse these processes, denoising the corrupted sequence back to the original data. Building on these foundations, subsequent models such as DiffusionBERT [58], LLaDA [44], and MD4 [35] introduce improvements in noise scheduling, scalability, and training objectives. Methods like MaskGIT [128] and FiLM [129], although originally proposed for vision or general infilling tasks, are methodologically aligned with mask-based diffusion, employing iterative generation with absorbing masks. These models have achieved performance competitive with, or even superior to, autoregressive models in language modeling, infilling, and reasoning tasks.

In addition to mask-based approaches, the uniform (categorical) transition process, also formalized in D3PM, corrupts sequences by progressively replacing tokens in the original data with tokens sampled uniformly from the vocabulary, rather than a single mask token. SEDD [34] extends score matching to discrete data via a score entropy loss, achieving state-of-the-art results and in some cases surpassing autoregressive baselines. RDM [130] introduces a reparameterized sampling framework to improve training and sampling efficiency. Furthermore, recent studies [131, 132] model discrete diffusion as a continuous-time Markov chain, advancing theoretical understanding and practical efficiency. Most recently, Discrete Flow Matching (DFM) [37] was proposed as a novel discrete flow paradigm for generative modeling of high-dimensional discrete data. Unlike flow matching and diffusion models designed for continuous domains, DFM introduces a general family of probability paths that interpolate between source and target distributions in discrete space, and provides a unified formula for sampling from these paths using learned posteriors such as probability denoisers and noise predictors. Empirically, DFM demonstrates that adopting a uniform (categorical) transition process, rather than an absorbing (mask-based) process, consistently leads to improved generative performance.

Recent scaling studies further demonstrate that, in addition to matching autoregressive models in perplexity and generation quality, discrete diffusion models have achieved strong performance on complex reasoning and planning tasks, underscoring their flexibility and potential as competitive alternatives for natural language generation and understanding [133-136, 44, 35]. Recent work [49] explores directly adapting pretrained autoregressive language models into non-autoregressive diffusion models via continual finetuning, enabling efficient knowledge transfer between paradigms. Building on this line, Dream 7B [45] further advances diffusion LMs by consistently outperforming previous diffusion models and matching the performance of top autoregressive models of similar size.

B More Comparison with State-of-the-arts

Qualitative Comparisons on Visual Generation. Figure 6 presents qualitative comparisons of visual generation results produced by three models: Janus [20], D-DiT [46], and our method, FUDOKI, across a diverse set of text prompts. Each row corresponds to a different prompt, covering scenarios such as animals in unusual environments, cartoon avatars, and objects with specific attributes. As shown in the figure, FUDOKI consistently produced images that more accurately captured the semantics of the prompts, demonstrating superior text-image alignment and higher visual fidelity.

Qualitative Comparisons on Visual Understanding. Figure 7 presents qualitative comparisons of visual understanding capabilities among Janus (AR) [20], D-DiT (mask-based discrete diffusion, MDD) [46], and our FUDOKI (discrete flow matching, DFM). The upper section shows selected intermediate outputs from each model’s answer generation process, illustrating their reasoning dynamics. The lower section presents additional visual question answering cases, where FUDOKI demonstrates higher reasoning accuracy and better alignment with ground truth answers, highlighting its superior ability to generate reliable and precise responses.

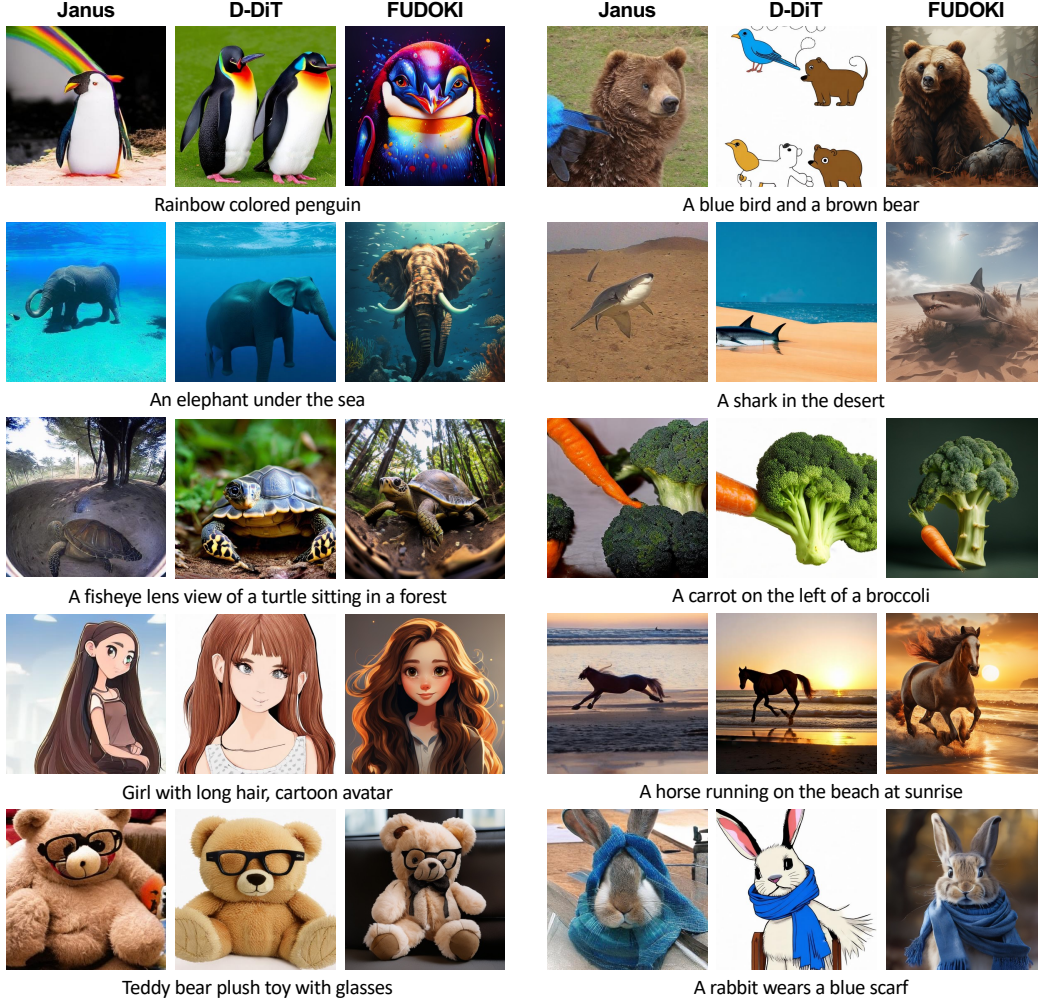


Figure 6: **Qualitative Comparisons on Visual Generation.** Comparison among Janus [20], D-DiT [46] and FUDOKI on various text prompts. The results demonstrate that our method (FUDOKI) achieved superior text-image alignment and aesthetics.

C Further Results

The Denoising Process of FUDOKI. Fig. 8 illustrates the iterative refinement process enabled by the discrete flow matching framework in FUDOKI, demonstrating its application to both generation and understanding tasks. The top panel visualizes how images are progressively denoised over iterations, transitioning smoothly from an initial noisy prior x_0 to the final high-fidelity image x_1 . Across diverse generation examples—ranging from animals to objects—the model incrementally sharpens semantic details and corrects spatial structure at each refinement step. The bottom panel depicts a similar iterative refinement for the understanding task, where the model extracts text from an image. Starting from a noisy token sequence, irrelevant or incorrect tokens are gradually replaced with accurate tokens (e.g., “Sara Lee”) as the model converges to the correct answer. The red arrows highlight token-level updates during each step, emphasizing the model’s ability to systematically and continuously correct errors and align predictions. This figure showcases how discrete flow matching enables fine-grained control and progressive improvement in both modalities by modeling transitions in discrete space, leading to more accurate and coherent outputs. More cases can be found in our project page: fudoki-dfm.github.io/fudoki/.

Maze Navigation. In this section, we train our proposed FUDOKI model on a novel task—maze navigation—which simultaneously requires understanding and generation capabilities. To this end,

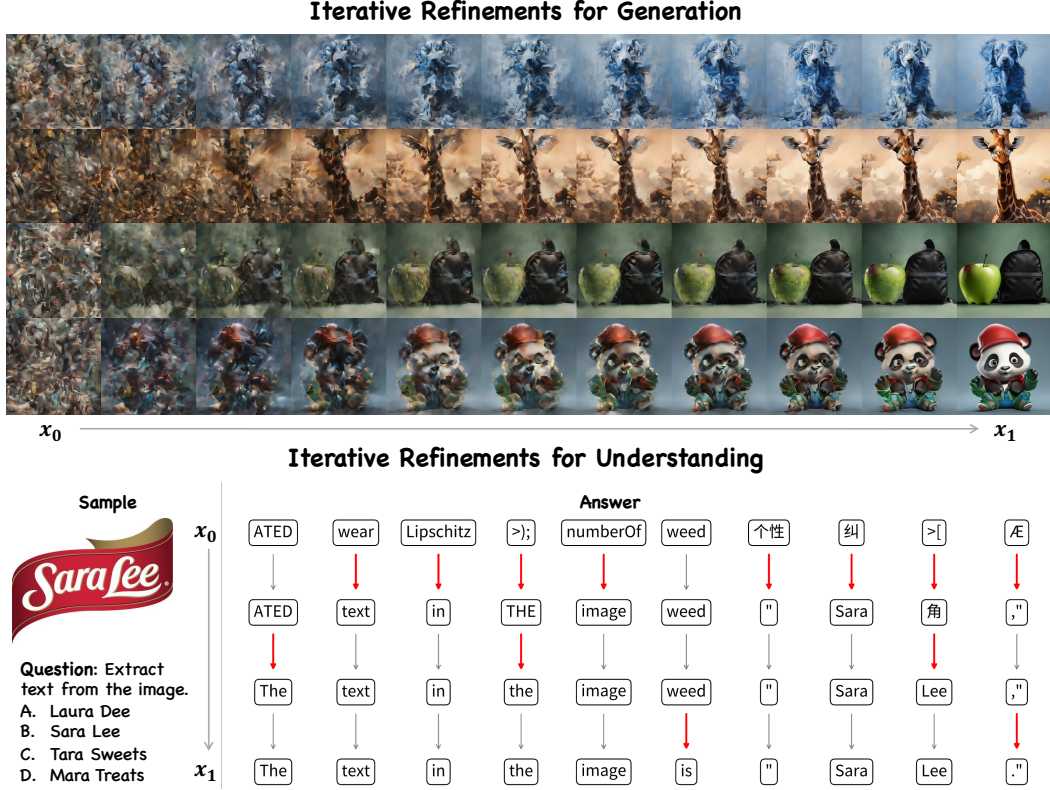


Figure 8: Visualization of the iterative refinement process enabled by discrete flow matching in FUDOKI, demonstrating denoising process for text-to-image generation and visual understanding tasks.

Table 5: Performance Comparisons on the MathVista Benchmark.

Method	Janus-1.5B	Janus-Pro-1B	FUDOKI
MathVista	32.4	35.1	38.6

spatial awareness. Furthermore, as shown in Fig. 10, FUDOKI is capable of completing the entire maze navigation sequence, moving from the initial position to the treasure step by step.

Results on the MathVista [137] Benchmark. We also evaluated our proposed FUDOKI on a more challenging mathematical reasoning benchmark, MathVista (testmini) [137]. As shown in Table 5, we find that FUDOKI achieved the best performance compared to AR-based models at the same scale. We attribute this improvement to FUDOKI’s discrete flow matching framework, which leverages bidirectional context modeling to facilitate complex reasoning.

D Dataset Collections

Our training set comprises a total of 12.62 million samples, divided into two main categories: Generation (8.76M, 69%) and Understanding (3.86M, 31%), as shown in Fig. 11. The Generation subset, which is entirely composed of in-house data, is constructed for text-to-image generation tasks. In contrast, the Understanding subset covers a diverse set of information extraction and comprehension tasks. This balanced and large-scale collection ensures comprehensive support for both generative and understanding capabilities.

Specifically, the public Understanding of data covers the following aspects:

- **General** (1506.8K, 40.6%): ShareGPT-4o (57.2K) [138], VSR (12.8K) [139], ALLaVA-Instruct (680.4K) [140], IconQA (29.9K) [141], LVIS-Instruct4V (10.0K) [142], ShareGPT4V (613.3K) [143], VIQuAE (18.5K) [144], RAVEN (0.3K) [145], Visual7W (14.4K) [146], In-house (70.0K)

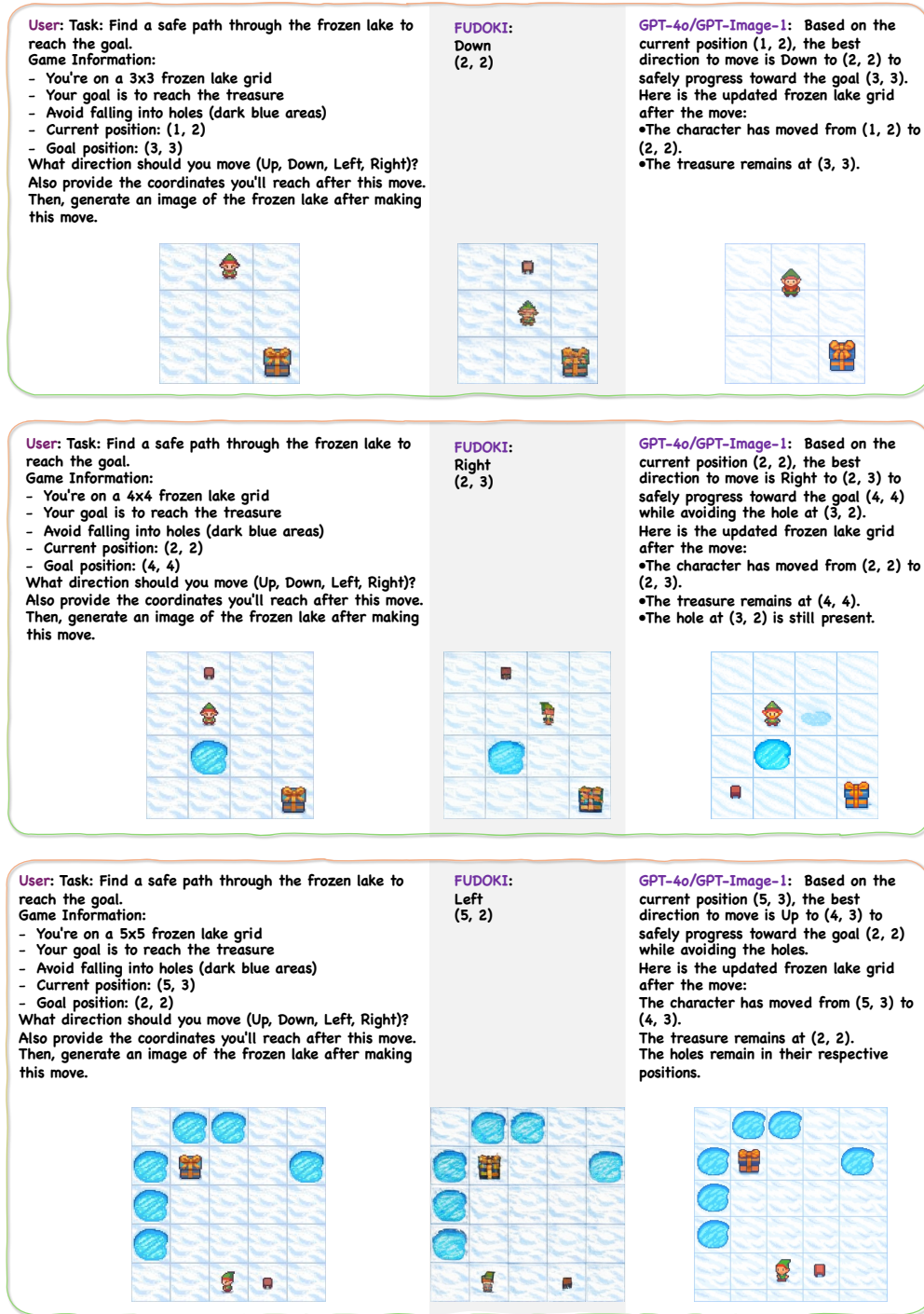


Figure 9: Comparison of FUDOKI and GPT-4o/GPT-Image-1 on frozen lake maze navigation tasks. GPT-4o/GPT-Image-1 offered well-reasoned textual outputs with safety and goal awareness but generated inconsistent visuals, even altering the maze (e.g., the third row). FUDOKI, by contrast, consistently produced valid directions and coherent visual updates aligned with task constraints, demonstrating stronger spatial consistency.

- **OCR** (428.0K, 11.5%): LLaVAR (59.3K) [61], SROIE (17.1K) [147], FUNSD (6.8K) [148], OCRVQA (80K) [149], MLHME-38K (30K) [150], Rendered Text (10.0K) [62], IIIT5K

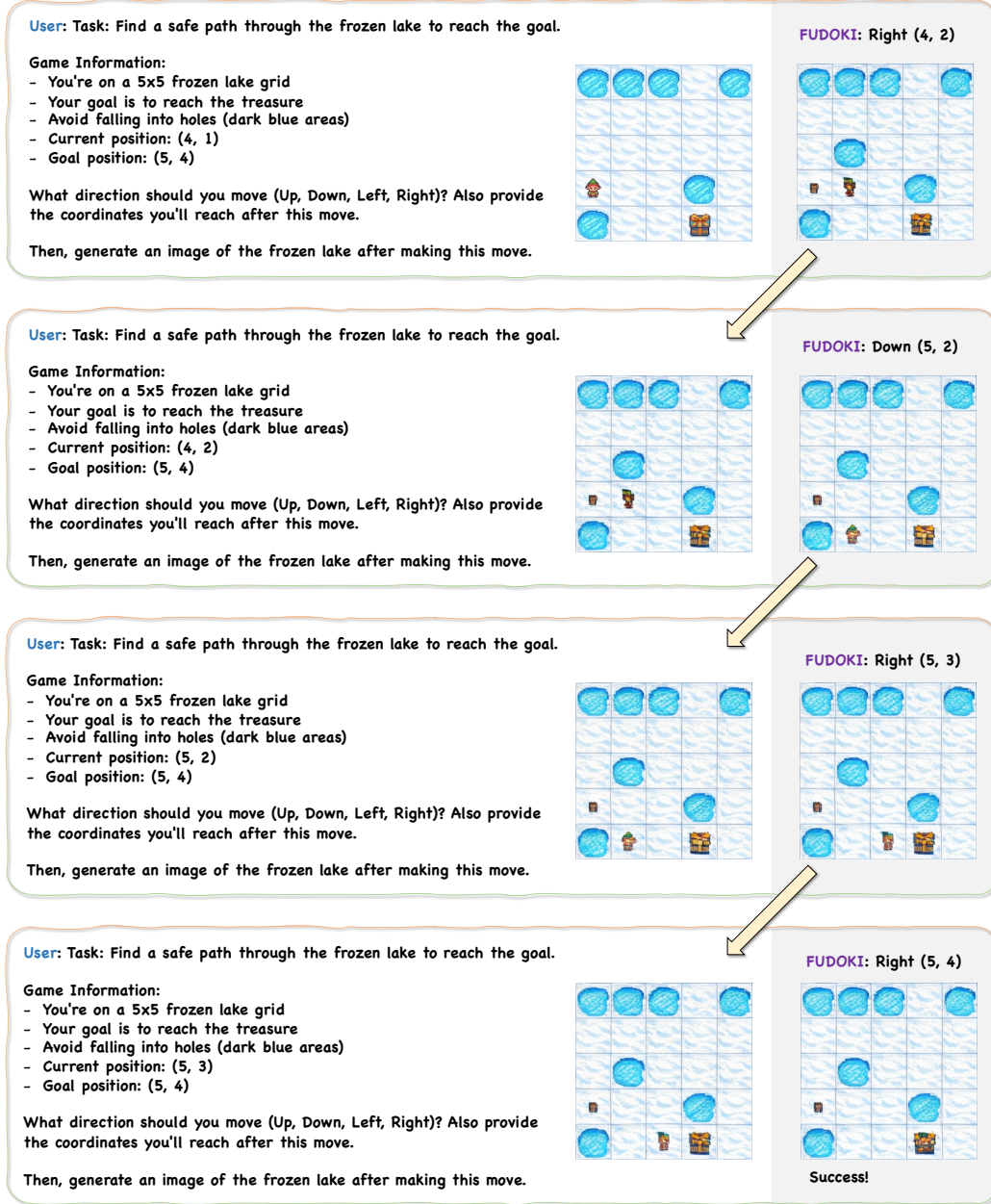


Figure 10: **FUDOKI successfully completed the full maze navigation task step by step.** Starting from the initial position at (4, 1), it sequentially selected safe moves—Right → Down → Right → Right—while avoiding holes and progressing toward the treasure at (5, 4). At each step, FUDOKI generated an updated image of the frozen lake, reflecting the character’s new position and preserving the environment’s structure, culminating in a successful arrival at the goal. Notably, in rows 2 through 4, the input images were taken directly from FUDOKI’s previous outputs, demonstrating the model’s ability to maintain coherent state tracking and visual continuity throughout the multistep decision-making process.

(6.0K) [151], HME100K (74.5K) [152], SynthDoG-EN (29.8K) [153], POIE (9.4K) [154], IAM (5.7K) [155], TextCaps (60.5K) [156], COCO-Text V2.0 (28.1K) [157], ChromeWriting (8.8K) [62], ORAND-CAR (2K) [158]

- **Document** (155.8K, 4.2%): DocVQA (122.4K) [63], FUNSD (6.8K) [148], Deepform (9.2K) [159], Kleister CharityAI (15.2K) [160], TAT-DQA (2.2K) [161]

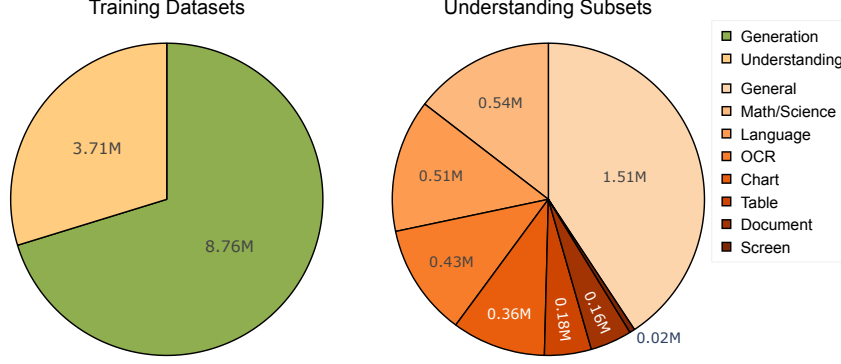


Figure 11: **Training Dataset Distribution.** The overall training data consists of 8.76M Generation samples (69%) and 3.86M Understanding samples (31%), as shown on the left. The right chart depicts the composition of the Understanding subset by category.

- **Table** (180.2K, 4.9%): TabFact (65.6K) [161], WikiTable (29.5K) [162], TabMWP (38.4K) [163], RoBUT WTQ (38.2K) [164], RoBUT SQA (8.5K) [164]
- **Chart** (362.6K, 9.8%): ChartQA (62.9K) [165], Chart2Text (27.0K) [64], PlotQA (10K) [166], DVQA (200K) [167], Infographic VQA (47.6K) [168], VisText (10.0K) [169], Diagram Image2Text (0.3K) [170], LRV Chart (1.8K) [171]
- **Screen** (24.6K, 0.7%): WebSRC (5.1K) [172], VisualMRC (19.5K) [65]
- **Math/Science** (544.9K, 14.7%): MAVIS (187.3K) [173], G-LLaVA (162.4K) [66], GeoQA+ (72.3K) [67], GeoMVerse (9.3K) [174], Geometry3K (3.0K) [175], MathVision (3.0K) [176], Cambrian Data Engine (50.8K) [177], Textbook QA (21.8K) [178], ScienceQA (19.2K) [179], AI2d (18.8K) [180]
- **Language** (510.2K, 13.7%): MathInstruct (81.5K) [181], Evol-Instruct (142.8K) [182], MathPlus (95.2K) [183], Magpie Pro (L3 MT) (50.0K) [68], ShareGPT4 (40.7K) [184], Magpie Pro (L3 ST) (50.0K) [68], Magpie Pro (Qwen2 ST) (50.0K) [68]

E Mathematical Formulations of Kinetic Optimal Velocity

To facilitate understanding, we use a simplified notation here and let \mathcal{T} denote the finite discrete state space, with elements $x, z \in \mathcal{T}$ (in the main paper, we have $x^i, z^i \in \mathcal{T}$). A probability path is a time-varying distribution $p_t(x)$, and a velocity field $u_t(x, z)$ describes mass transport between states over time. In this way, we have the *Continuity Equation* as follows.

$$\dot{p}_t(x) + \text{div}_x(j_t) = 0, \quad \forall x \in \mathcal{T}$$

with the discrete divergence given by $\text{div}_x(j_t) = \sum_{z \neq x} j_t(z, x) - \sum_{z \neq x} j_t(x, z)$ and $j_t(x, z)$ is the flux, defined by $j_t(x, z) = u_t(x, z) p_t(z)$, which represents the flow of probability mass from z to

x . In this way, the velocity can be obtained by $u_t(x, z) = \begin{cases} \frac{j_t(x, z)}{p_t(z)} & \text{if } p_t(z) > 0 \\ 0 & \text{otherwise} \end{cases}$ when $x \neq z$ and $u_t(z, z) = -\sum_{x \neq z} u_t(x, z)$ to ensure the rate condition in Eq. 2. With such notations, we expect to minimize the kinetic energy during the flow process, namely,

$$\min_{p_t, j_t} \int_0^1 \sum_{x \neq z} w_t(x, z) \frac{j_t(x, z)^2}{p_t(z)} dt$$

subject to:

- Continuity Equation: $\text{div}_x(j_t) = -\dot{p}_t(x)$
- Non-negativity of the flux: $j_t(x, z) \geq 0 \quad \forall x \neq z$
- Boundary conditions: $p_0 = p, \quad p_1 = q$

Here, $w_t(x, z) > 0$ is a problem-specific weight controlling the "cost" of mass moving from z to x . As evidenced in [38], when p_t is given and let $w_t(x, z) = 1/p_t(x)$, the kinetic optimal solution can be obtained via $j_t^*(x, z) = [p_t(z)\dot{p}_t(x) - \dot{p}_t(z)p_t(x)]_+ \quad \forall x \neq z$. In this way, if we apply this kinetic optimal $j_t^*(x, z)$ for the probability path in Eq. 4, we can obtain the velocity defined in Eq. 5.

F Limitations and Broader Impacts

Limitations. Despite its promising results, FUDOKI also presents several limitations that warrant further investigation. First, despite the advantages of discrete flow matching—such as being agnostic to token order and compatible with bidirectional Transformers—the current implementation requires the sequence length to be fixed prior to sampling. This constraint limits flexibility in generation and makes dynamic-length outputs challenging. A promising direction for future work is to extend the sampling scheme to support variable-length generation, which would broaden the applicability of the model across open-ended tasks and enhance the flexibility on the computational cost during inference. Besides, as shown in Fig. 12, while FUDOKI shows strong performance, it still faces challenges under certain scenarios, such as performing text-to-image generation given complex prompts or prompts involving rendering specific texts in images, as well as performing visual understanding tasks that demand expert-level reasoning and domain-specific knowledge.

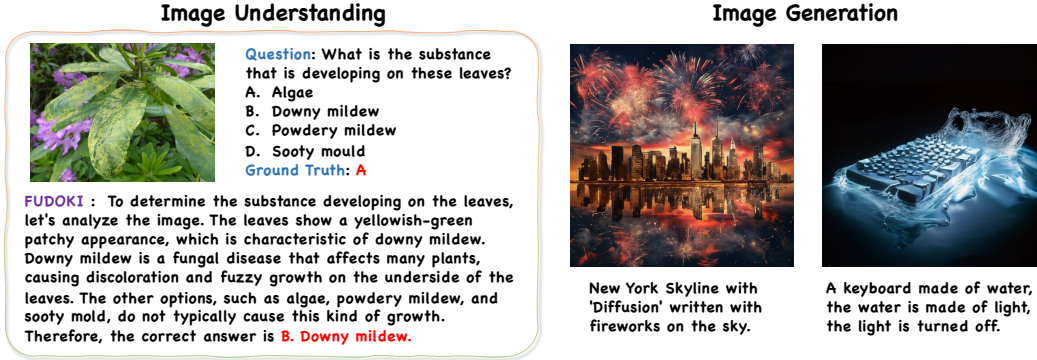


Figure 12: Examples of failed cases on visual understanding and generation. While FUDOKI demonstrated strong performance, it still struggled with harder tasks—such as generating images from complex prompts involving specific texts, and understanding visuals that require expert-level knowledge.

Broader Impacts. FUDOKI introduces a novel paradigm for unified multimodal modeling that departs from the long-dominant autoregressive approach, potentially redefining how future multimodal systems are designed. By leveraging discrete flow matching with metric-induced probability paths, FUDOKI enables controllable and interpretable generation processes, which could prove valuable in critical applications such as education, embodied AI, and autonomous driving. Its iterative, self-correcting refinement process aligns well with human reasoning patterns and may support safer, more reliable AI agents in domains requiring high precision, such as medicine and law. Furthermore, FUDOKI’s unified architecture for both understanding and generation fosters more integrated, general-purpose agents—an important step toward realizing practical artificial general intelligence (AGI). However, as with any generative technology, ethical considerations around bias, misuse, and content safety must be carefully addressed as adoption scales.

References

- [1] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- [2] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, et al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024.
- [4] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, et al. Internlm2 technical report, 2024.
- [5] OpenAI. Chatgpt. <https://chat.openai.com/>, 2023.
- [6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [7] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [8] 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.
- [9] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [10] 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.
- [11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [12] 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.
- [13] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.
- [14] Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206, 2024.
- [15] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [16] 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.
- [17] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

- [18] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yuezhe Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need, 2024.
- [19] 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.
- [20] 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. *ArXiv*, abs/2410.13848, 2024.
- [21] 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.
- [22] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024.
- [23] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.
- [24] Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models. *arXiv preprint arXiv:2503.08686*, 2025.
- [25] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.
- [26] 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. *ArXiv*, abs/2501.17811, 2025.
- [27] 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.
- [28] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [29] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023.
- [30] Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. *ArXiv*, abs/2403.06963, 2024.
- [31] Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. *ArXiv*, abs/2410.14157, 2024.
- [32] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *ArXiv*, abs/2310.01798, 2023.
- [33] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

- [34] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32819–32848, 2024.
- [35] Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167, 2024.
- [36] Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.
- [37] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.
- [38] Neta Shaul, Itai Gat, Marton Havasi, Daniel Severo, Anuroop Sriram, Peter Holderrieth, Brian Karrer, Yaron Lipman, and Ricky T. Q. Chen. Flow matching with general discrete paths: A kinetic-optimal perspective. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [39] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *International Conference on Machine Learning*, pages 5453–5512. PMLR, 2024.
- [40] Mercury coder, 2025. URL <https://www.inceptionlabs.ai/news>
- [41] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- [42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [43] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.
- [44] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.
- [45] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025. URL <https://hkunlp.github.io/blog/2025/dream>
- [46] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. *arXiv preprint arXiv:2501.00289*, 2024.
- [47] Minghui Hu, Chuanxia Zheng, Heliang Zheng, Tat-Jen Cham, Chaoyue Wang, Zuopeng Yang, Dacheng Tao, and Ponnuthurai N Suganthan. Unified discrete diffusion for simultaneous vision-language generation. *arXiv preprint arXiv:2211.14842*, 2022.
- [48] Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.
- [49] Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*, 2024.
- [50] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.
- [51] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025.

- [52] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025.
- [53] 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. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [54] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [55] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin CHEN, Chengru Song, dai meng, Di ZHANG, Wenwu Ou, Kun Gai, and Yadong MU. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *The Twelfth International Conference on Learning Representations*, 2024.
- [56] 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.
- [57] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [58] Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuan-Jing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4521–4534, 2023.
- [59] 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.
- [60] 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.
- [61] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavir: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.
- [62] Chris Wendler. wendlerc/renderedtext, 2023.
- [63] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.
- [64] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model, 2020. URL <https://arxiv.org/abs/2010.09142>.
- [65] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, 2021.
- [66] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjuan Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model, 2023. URL <https://arxiv.org/abs/2312.11370>.
- [67] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning, 2022. URL <https://arxiv.org/abs/2105.14517>.

- [68] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *ArXiv*, abs/2406.08464, 2024. URL <https://api.semanticscholar.org/CorpusID:270391432>.
- [69] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [70] 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.
- [71] 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.
- [72] 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.
- [73] 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.
- [74] 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. *arXiv preprint arXiv:2412.03069*, 2024.
- [75] 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.
- [76] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.
- [77] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. *arXiv preprint arXiv:2406.18583*, 2024.
- [78] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.
- [79] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.
- [80] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.
- [81] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.
- [82] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

- [83] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-phi: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024.
- [84] 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.
- [85] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [86] Hugo Laurençon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, and et al. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023. URL <https://huggingface.co/blog/idefics>.
- [87] 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.
- [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. *arXiv preprint arXiv:2412.14164*, 2024.
- [89] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [90] 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.
- [91] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [92] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [93] 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.
- [94] 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? *arXiv preprint arXiv:2307.06281*, 2023.
- [95] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- [96] 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.
- [97] 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.
- [98] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024.

- [99] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.
- [100] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024.
- [101] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.
- [102] 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.
- [103] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.
- [104] Kaihang Pan, Wang Lin, Zhongqi Yue, Tenglong Ao, Liyu Jia, Wei Zhao, Juncheng Li, Siliang Tang, and Hanwang Zhang. Generative multimodal pretraining with discrete diffusion timestep tokens. *arXiv preprint arXiv:2504.14666*, 2025.
- [105] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jian-jian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.
- [106] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.
- [107] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.
- [108] 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.
- [109] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.
- [110] Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR 2023 Conference*, 2023.
- [111] Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23263–23274, 2023.
- [112] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025.

- [113] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>.
- [114] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duoju Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xincheng Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025.
- [115] Alexander H Liu, Matt Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu. Generative pre-training for speech with flow matching. *arXiv preprint arXiv:2310.16338*, 2023.
- [116] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.
- [117] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.
- [118] Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.
- [119] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *International Conference on Machine Learning*, pages 22277–22303. PMLR, 2024.
- [120] Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se (3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2023.
- [121] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0 : A vision-language-action flow model for general robot control, 2024.
- [122] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.
- [123] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [124] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343, 2022.
- [125] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.

- [126] Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715, 2023.
- [127] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.
- [128] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.
- [129] Tianxiao Shen, Hao Peng, Ruoqi Shen, Yao Fu, Zaid Harchaoui, and Yejin Choi. Film: Fill-in language models for any-order generation. *arXiv preprint arXiv:2310.09930*, 2023.
- [130] Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. In *First Conference on Language Modeling*, 2024.
- [131] Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.
- [132] Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.
- [133] Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text. *arXiv preprint arXiv:2410.18514*, 2024.
- [134] Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, et al. Diffusion of thought: Chain-of-thought reasoning in diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [135] Jiacheng Ye, Zhenyu Wu, Jiahui Gao, Zhiyong Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Implicit search via discrete diffusion: A study on chess. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [136] Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [137] 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. In *The Twelfth International Conference on Learning Representations*.
- [138] Shanghai AI Laboratory. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2023.
- [139] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.
- [140] 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 lite vision-language models, 2024.
- [141] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS*, 2021.
- [142] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

- [143] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [144] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno, and Jesús Lovón Melgarejo. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’22*, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3477495.3531753.
- [145] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, 2019.
- [146] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [147] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019. doi: 10.1109/icdar.2019.00244.
- [148] Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*, 2019.
- [149] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.
- [150] Mlhme-38k, 2025. URL <https://ai.100tal.com/icdar>.
- [151] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.
- [152] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. *arXiv preprint arXiv:2203.01601*, 2022.
- [153] 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 (ECCV)*, 2022.
- [154] Jianfeng Kuang, Wei Hua, Dingkan Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pages 36–53. Springer, 2023.
- [155] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5: 39–46, 2002.
- [156] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020.
- [157] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.
- [158] Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M. Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S. Oliveira. Proceedings of 2014 14th International Conference on Frontiers in Handwriting Recognition, pages 779–784, 2014. doi: 10.1109/ICFHR.2014.136.
- [159] Deepform, 2025. URL https://wandb.ai/stacey/deepform_v1/reports/DeepForm-Understand-Structured-Documents-at-Scale--VmlldzoyODQ3Njg.

- [160] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021.
- [161] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022.
- [162] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.
- [163] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023.
- [164] Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. Robut: A systematic study of table qa robustness against human-annotated adversarial perturbations. *arXiv preprint arXiv:2306.14321*, 2023.
- [165] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022.
- [166] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020.
- [167] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.
- [168] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.
- [169] Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning, 2023. URL <https://arxiv.org/abs/2307.05356>
- [170] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [171] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.
- [172] 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.
- [173] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, and Hongsheng Li. Mavis: Mathematical visual instruction tuning, 2024. URL <https://arxiv.org/abs/2407.08739>
- [174] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.
- [175] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning, 2021. URL <https://arxiv.org/abs/2105.04165>

- [176] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.
- [177] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.
- [178] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.
- [179] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [180] 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.
- [181] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.
- [182] Chandeeppa Dissanayake, Lahiru Lowe, Sachith Gunasekara, and Yasiru Ratnayake. Open-bezoar: Small, cost-effective and open models trained on mixes of instruction data. *arXiv preprint arXiv:2404.12195*, 2024.
- [183] Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhui Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660, 2024.
- [184] Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*, 2024.