



FUDOKI: Discrete Flow-based Unified Understanding and Generation via Kinetic-Optimal Velocities

Jin Wang^{*,1} Yao Lai^{*,1} Aoxue Li² Shifeng Zhang² Jiacheng Sun² Ning Kang² Chengyue Wu¹
Zhenguo Li^{†,2} Ping Luo^{†,1}

¹The University of Hong Kong ²Huawei Noah's Ark Lab

<https://fudoki-hku.github.io/>

Abstract

The rapid progress of large language models (LLMs) has catalyzed the emergence of multimodal large language models (MLLMs) that unify visual understanding and image generation within a single framework. However, most existing MLLMs rely on autoregressive (AR) architectures, which impose inherent limitations on future development, such as the raster-scan order in image generation and restricted reasoning abilities in causal context modeling. In this work, we challenge the dominance of AR-based approaches by introducing FUDOKI, a unified multimodal model purely based on discrete flow matching, as an alternative to conventional AR paradigms. By leveraging metric-induced probability paths with kinetic optimal velocities, our framework goes beyond the previous masking-based corruption process, enabling iterative refinement with self-correction capability and richer bidirectional context integration during generation. To mitigate the high cost of training from scratch, we initialize FUDOKI from pre-trained AR-based MLLMs and adaptively transition to the discrete flow matching paradigm. Experimental results show that FUDOKI achieves performance comparable to state-of-the-art AR-based MLLMs across both visual understanding and image generation tasks, highlighting its potential as a foundation for next-generation unified multimodal models. Furthermore, we show that applying test-time scaling techniques to FUDOKI yields significant performance gains, further underscoring its promise for future enhancement through reinforcement learning.

1. Introduction

Driven by the rapid progress of large language models (LLMs) [1–5], a new wave of large-scale multimodal models has emerged, delivering remarkable advances in the two fundamental pillars of artificial general intelligence (AGI): understanding [6–10] and generation [11–15]. Building on this momentum, a growing body of work [16–21] seeks to unify perception and synthesis within a single framework, introducing versatile multimodal large language models (MLLMs) that seamlessly integrate visual understanding with image generation.

* Equal Contribution

† Correspondence to: Zhenguo Li <li.zhenguo@huawei.com> and Ping Luo <pluo@cs.hku.hk>.

In prior research, most MLLMs adopt the autoregressive (AR) architecture of standard LLMs, processing multimodal tokens sequentially from left to right for both understanding and generation tasks [22, 23]. While these MLLMs deliver strong performance across many multimodal tasks, their inherent AR design’s limitations have become increasingly apparent as shown in recent studies, such as weaker performance in complex reasoning [24–26], challenges in future planning [27], and difficulties with self-correction [28]. These shortcomings are particularly critical for emerging domains such as embodied AI and autonomous agents, where complex reasoning and deep contextual understanding are essential. This prompts a fundamental question for the future of AGI development: *what architectural paradigm could define the next generation of MLLMs?*

To this end, discrete-space generative flow and diffusion models have gained attention as a promising alternative for generative modeling. These models have seen success in the domain of text generation [29–34], protein design [35], image synthesis [33, 34], and code generation [33, 36]. Unlike sequential autoregressive models, these models usually begin with a fully corrupted sequence and iteratively denoise the entire sequence in parallel, which allows richer integration of information from both directions to enhance prolonged reasoning. Moreover, these models enable flexible and controllable generation through their inherent iterative refinement process, while offering the potential for accelerated sampling via novel training designs [37–39]. Recent studies like LLaDA [40] and Dream [41] have also scaled discrete diffusion models to 7B parameters, further highlighting their growing potential to overcome the fundamental limitations of autoregressive approaches.

To advance the application of discrete generative flow modeling and challenge the dominance of the AR-based paradigm in MLLMs, we present FUDOKI, a unified multimodal model purely based on *discrete flow matching*. Different from previous diffusion-based unified multimodal models [42–44] focusing solely on the case of masking as a corruption process, we adopt the novel framework of discrete flow matching [33, 34], which substantially expanded the design space of discrete-space generative models by enabling metric-induced probability paths with kinetic optimal velocities. This design enables better performance than masked construction [34] and allows models to continuously self-correct their responses during the iterative refinement process. Moreover, to mitigate the high training cost of training large discrete flow matching models for multimodal tasks, we leverage the pre-trained AR-based MLLM [20] as the initialization and adaptively transfer it to the discrete flow matching paradigm [45].

The contributions of this paper can be summarized as follows: 1) We introduce FUDOKI³, the first general-purpose unified multimodal model built entirely on discrete flow matching. Unlike traditional approaches that rely on masking-based corruption, FUDOKI leverages a metric-induced probability path with kinetically optimal velocities, expanding the design space of discrete multimodal modeling and offering advantages during inference; 2) Through extensive experiments, we show that FUDOKI achieves competitive performance on both visual understanding and text-to-image generation tasks, rivaling autoregressive-based MLLMs; 3) We apply test-time inference scaling techniques to FUDOKI inspired by [46], which yield substantial improvements across visual generation and understanding benchmarks. This suggests strong potential for future enhancement of FUDOKI via reinforcement learning [1, 47]. We believe that FUDOKI provides a compelling foundation for the development of next-generation unified multimodal models.

³ 風土記 (*FUDOKI*) is a Japanese term referring to ancient records that comprehensively document and integrate the culture, geography, and traditions of different regions. We name our model *FUDOKI* to highlight its unified ability to both understand and generate multimodal information, such as interpreting and generating diverse images, mirroring how the original *FUDOKI* integrates and presents multifaceted knowledge.



Figure 1 | Qualitative Results of Visual Generation and Understanding Capabilities of FUDOKI. FUDOKI is designed based on the discrete flow matching for both visual and textual modalities, capable of performing understanding and generation simultaneously under one unified paradigm.

2. Preliminary: Discrete Flow Matching

In this section, we present key concepts and notations in discrete flow matching [33] to facilitate understanding in the following sections. Generally speaking, the objective of discrete flow matching is to approximate the target underlying data distribution $q(x)$ from the source known distribution $p(x)$, where $x = (x^1, x^2, \dots, x^D)$ belongs to the discrete space $\mathcal{S} = \mathcal{T}^D$, where D is

the number of discrete variables and $\mathcal{T} = [K] = \{1, 2, \dots, K\}$ represents a finite set of possible discrete values.

Probability Paths. Given a *source distribution* $p(x)$ and a *target distribution* $q(x)$ defined over a finite state space \mathcal{S} , discrete flow matching defines a family of time-indexed probability distributions $\{p_t(x)\}_{t \in [0,1]}$ to describe a smooth transformation from p to q , referred to as *probability paths*. Each $p_t(x)$ is constructed as: $p_t(x) := \sum_{x_1 \in \mathcal{S}} p_t(x | x_1) q(x_1)$, where the conditional distribution is factorized across dimensions, namely $p_t(x | x_1) := \prod_{i=1}^D p_t(x^i | x_1^i)$. Here, each $p_t(x^i | x_1^i)$ defines a univariate interpolation between a base distribution $p(x^i)$ and a point mass $\delta_{x_1^i}(x^i)$, i.e., $\delta_{x_1^i}(x^i) = 1$ if $x^i = x_1^i$ else 0. A common design for such interpolations is the *mixture path*, defined via a time-dependent scheduler $\kappa_t(x_1^i) \in [0, 1]$:

$$p_t(x^i | x_1^i) = (1 - \kappa_t(x_1^i))p(x^i) + \kappa_t(x_1^i)\delta_{x_1^i}(x^i), \quad (1)$$

where $\kappa_0(\cdot) = 0$ and $\kappa_1(\cdot) = 1$. This class of paths recovers the masked data construction when $p(x^i) = \delta_m(x^i)$ with m denoting the *mask token*, which are widely used in previous studies [31, 32].

Probability Velocities. To simulate the generative process that evolves along the prescribed path $\{p_t(x)\}_{t \in [0,1]}$, we consider a continuous-time Markov chain (CTMC) $\{x_t\}_{t \in [0,1]}$ over the discrete space \mathcal{S} , such that: $x_t \sim p_t$. Specifically, we describe this CTMC via a *probability velocity* $u_t^i(\cdot, x_t)$ (also known as the rate matrix), describing the rate of probability change of x_t in its i -th token. Reminiscent of the velocity field in the continuous Flow Matching [38, 37], discrete flow matching features the following definition:

Definition 1. A probability velocity u_t is said to generate the probability path p_t if, for all $t \in [0, 1)$ and for any sample $x_t \sim p_t$, the updated sample $x_{t+h}^i \sim \delta_{x_t^i}(\cdot) + h u_t^i(\cdot, x_t)$ for each coordinate i satisfies the condition that $x_{t+h} \sim p_{t+h} + o(h)$ ⁴ as $h \rightarrow 0$.

Besides, the probability velocity u_t should satisfy the following *rate condition*:

$$\sum_{x^i \in [K]} u_t^i(x^i, z) = 0, \quad \text{and} \quad u_t^i(x^i, z) \geq 0 \quad \forall i \in [D], x^i \neq z^i, \quad (2)$$

such that the updated x_{t+h}^i can be sampled from a valid probability distribution. Further, previous studies [33, 35] also demonstrate the *Continuity Equation* (also known as the Kolmogorov forward equation) in discrete flow matching, which describes the state probability rate $\dot{p}_t(x)$, $x \in \mathcal{S}$ by:

$$\dot{p}_t(x) + \text{div}_x(p_t u_t) = 0. \quad (3)$$

where $\text{div}_x(p_t u_t) = \sum_{z \in \mathcal{S}} \sum_{i=1}^D \delta_x(z^i) [p_t(x) u_t^i(z^i, x) - p_t(z) u_t^i(x^i, z)]$, measuring the total outgoing flux $x \rightarrow z$ minus the total incoming flux $z \rightarrow x$ for state $x \in \mathcal{S}$. Here $\delta_x(z^i) = \prod_{j \neq i} \delta_{x^j}(z^j)$, which indicates that we only consider x and z when they only differ in the i -th coordinate for calculating the flux [33, 30]. Intuitively, Eq. 3 expresses that the rate of probability at x is equal to the final remaining probability flux $p_t u_t$ at x . Previous studies [33, 35] have shown that if the Continuity Equation is satisfied, then u_t is said to generate the probability path p_t as in Definition 1.

3. FUDOKI: A Multimodal Model Purely Based on Discrete Flow Matching

This section introduces FUDOKI, a new multimodal architecture that unifies vision and language through the novel lens of discrete flow matching. By adopting this framework, FUDOKI enables an integrated approach to both perception and generation across visual and textual modalities.

⁴ $o(h)$ refers to a function that vanishes at a faster rate than h as $h \rightarrow 0$, i.e., $\lim_{h \rightarrow 0} \frac{o(h)}{h} = 0$.

3.1. Metric-induced Probability Paths with Kinetic Optimal Velocities

Based on the recent theoretical advancement of discrete flow matching [34], we adopt a more general probability path for FUDOKI, instead of the commonly used mask-based mixture paths [33, 32, 31, 42, 41]. Specifically, we consider the probability paths induced by discrete metrics. Given a distance function $d : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}_{\geq 0}$ satisfying $d(x^i, x_1^i) = 0$ if and only if $x^i = x_1^i$, we define a path of conditional distributions via:

$$p_t(x^i | x_1^i) = \text{softmax}(-\beta_t \cdot d(x^i, x_1^i)), \quad (4)$$

where $\beta_t : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$ is a monotonic schedule with boundary values $\beta_0 = 0, \beta_1 = \infty$. At $t = 0$, this yields a uniform distribution, and as $t \rightarrow 1$, the distribution converges to a delta function at x_1^i . Compared to the previous mask-based probability path (*i.e.*, Eq. 1), this metric-induced probability path defines a more semantically meaningful transformation, allowing the probabilities of tokens similar to x_1^i to also increase as $t \rightarrow 1$, when setting $d(\cdot, \cdot)$ to measure token embedding distances.

After defining the prescribed metric-induced probability path, we then obtain the probability velocities via minimizing the kinetic energy [34]. In other words, it is expected to minimize the magnitude of flux $p_t u_t$ for probability velocities to obtain a smooth transformation along the probability path. Meanwhile, the obtained velocities should also satisfy several conditions, including the Continuity Equation (*i.e.*, Eq. 3), the non-negativity of the flux between different states (*i.e.*, Eq. 2), and the boundary conditions for p and q . We leave the detailed mathematical formulations in the appendix. In this way, the kinetic optimal velocity for Eq. 4 can be formulated as follows [34],

$$u_t^i(x^i, z | x_1) = p_t(x^i | x_1^i) \dot{\beta}_t [d(z^i, x_1^i) - d(x^i, x_1^i)]_+ \quad (5)$$

where $[\cdot]_+ = \max\{\cdot, 0\}$ is the ReLU operator and $\dot{\beta}_t$ is the derivative of β_t w.r.t t . Intuitively, for the i -th coordinate $z^i \in \mathcal{T}$, this velocity ensures that probability mass flows from state z^i to state x^i only when x^i lies closer to x_1^i than z^i does, *i.e.*, $d(x^i, x_1^i) < d(z^i, x_1^i)$. As a result, the flow monotonically progresses toward x_1^i . After introducing the mathematical foundation of discrete flow matching, we now dive into FUDOKI’s model structure details.

3.2. Architecture Overview

As shown in Fig. 2(e), FUDOKI is based on the Janus-1.5B [20] architecture, with minor adaptations to support unified vision-language discrete flow modeling. Specifically, to facilitate effective learning and accelerate convergence, 1) we adopt a full attention mask instead of the standard causal mask to allow all tokens to attend to each other, which helps the model better capture global context; 2) we apply a shifting operation [45] to the output logits by one position, so that our model can inherit the next-token prediction capabilities of AR-based MLLMs as much as possible; 3) unlike continuous diffusion models [53, 12], we do not incorporate additional time embedding layers in the model to explicitly indicate the noise level in the corrupted input. Following the intuition of mask-based discrete diffusion models [45, 54], we observe that our discrete generative model can also implicitly infer the timesteps from the corrupted input along our defined metric-induced probability path (*i.e.*, Eq. 4), resulting in faster adaptation in experiments. The rest of the architecture remains identical to Janus-1.5B. For the text modality, we use the tokenizer with a vocabulary size of 102,400. For images, we decouple the processing paths for understanding and generation. The semantic encoder SigLIP [55] extracts high-dimensional features for image understanding, which are reshaped and mapped into the LLM input space via an adaptor. For image generation, we follow LlamaGen [56], employing a pixel encoder

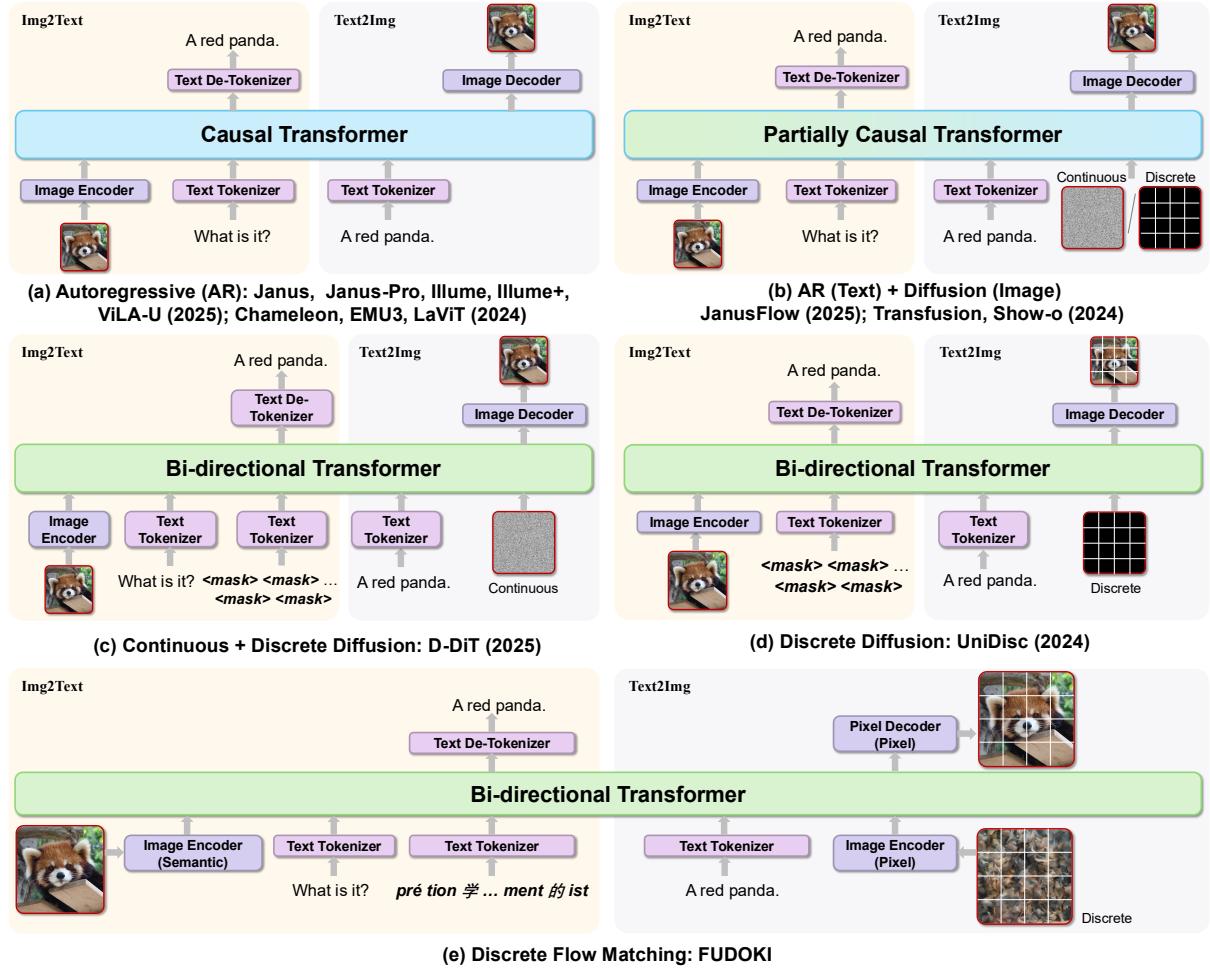


Figure 2 | Comparison of Model Architectures in Unified Multimodal Models. (a) AR-based models [20, 22, 21, 48–50, 18, 51] perform multimodal tasks via sequential token generation under strictly causal context modeling. (b) Hybrid AR+Diffusion models, such as Transfusion [19] and Show-o [52], integrate AR for text and diffusion models for images, enabling improved visual generation quality. (c-d) Diffusion-based models: D-DiT [42] applies mask-based discrete diffusion to text and continuous diffusion to images, while UniDisc [44] employs mask-based discrete diffusion for both modalities. (e) FUDOKI adopts a unified discrete flow matching framework for both modalities, leveraging a metric-induced probability path to enhance performance in understanding and generation tasks. The inference advantages of FUDOKI over mask-based discrete diffusion modeling used in (c-d) are shown in Fig. 3.

and decoder to convert images into discrete tokens, with the image token vocabulary size set to 16,384. Each image token embedding is further transformed into an input feature via a generation adaptor before being fed into the LLM. At the output stage, we use two output heads, a text head and an image head, which convert the transformer outputs into discrete categorical distributions. The appropriate head is selected depending on the target modality during inference. Comparisons with previous AR-based and diffusion-based MLLMs are shown in Fig. 2.

3.3. Training

We follow the discrete flow matching framework [30] for model training. Our model is initialized from the pretrained weights of Janus-1.5B [20] and further adapted to our collected dataset, which contains both text-to-image (generation) and image-to-text (understanding) data.

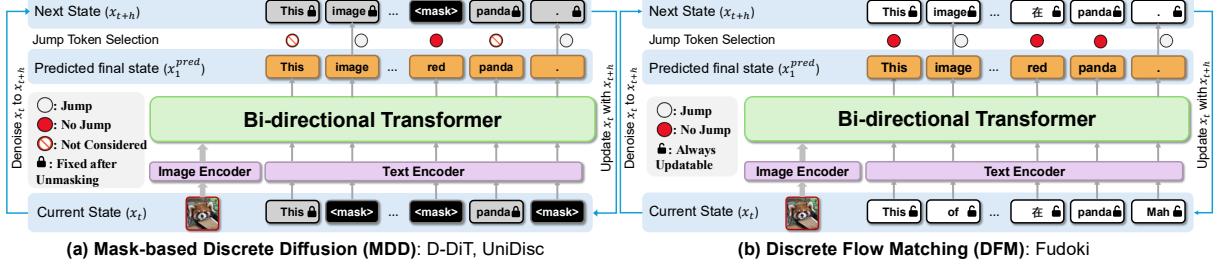


Figure 3 | Inference Comparisons between (a) Mask-Based Discrete Diffusion Models and (b) Discrete Flow Matching-Based FUDOKI. In mask-based discrete diffusion models, once a token is unmasked, it typically cannot be modified again, which hinders self-correction. In contrast, our proposed FUDOKI allows its responses to be continuously updated during inference, enabling potential corrections.

Specifically, we divide the training of FUDOKI into two stages: 1) The main goal of the first stage is to quickly relearn the AR-based LLM such that it can effortlessly support the discrete flow matching paradigm. To this end, we only fine-tune the parameters of the transformer while keeping other parts of the model frozen, including the semantic encoders and embedding adaptors. This can help accelerate convergence and stabilize our training; 2) After the first stage, we further fine-tune the whole model to enhance its overall performance on understanding and generation based on discrete flow matching.

Specifically, in each training stage, the ground-truth target x_1 is drawn from the data distribution $q(\cdot)$, where the condition is either a text prompt (for T2I) or an image-question pair (for I2T). The target x_1 is the image token sequence in the T2I setting and the textual token sequence in the I2T setting. At each training step, a time $t \in [0, 1]$ is uniformly sampled, and a noised sequence x_t is sampled according to the defined probability path $p_t(\cdot | x_1)$ in Eq. 4. We set the distance function $d(\cdot, \cdot)$ to measure the L2-distances between normalized token embeddings, which helps increase the probability of sampling tokens whose embeddings are close to the corresponding ground-truth token x_1^i in the embedding space, thereby making the corruption process more semantically meaningful and facilitating learning. The model then receives x_t as input and predicts x_1 , outputting per-token logits for each position. The training loss is defined as the expected cross-entropy between the ground-truth sequence x_1 and the model’s predicted distribution:

$$\mathcal{L}_{CE}(\theta) = \mathbb{E}_{t \sim U[0,1], x_1 \sim q(\cdot), x_t \sim p_t(\cdot | x_1)} \left[- \sum_{i=1}^D \log p_{1|t}^\theta(x_1^i | x_t) \right] \quad (6)$$

where $p_{1|t}^\theta(\cdot | x_t)$ denotes the model’s predicted categorical distribution for the i -th position, parameterized by θ , given input x_t .

3.4. Inference

During inference, we apply an Euler solver for more robust sampling as suggested in [34]. This solver simulates the continuous-time Markov chain (CTMC) process $(x_t)_{0 \leq t \leq 1}$. Given that $x_t \sim p_t$, the solver updates the i -th coordinate from time t to $t + h$ using the following procedure:

- Sample $x_1^i \sim p_{1|t}^\theta(\cdot | x_t)$ from our model;
- Compute the total conditional transition rate $\lambda^i = \sum_{x^i \neq x_t^i} u_t^i(x^i, x_t^i | x_1^i)$ (see Eq. 5);
- Draw a uniform random variable $Z_{\text{change}}^i \sim U[0, 1]$;

- Sample x_{t+h}^i as follows: if $Z_{\text{change}}^i \leq 1 - e^{-h\lambda^i}$, sample x_{t+h}^i from $\frac{u_t^i(\cdot, x_t^i | x_1^i)}{\lambda^i}(1 - \delta_{x_t^i}(\cdot))$; otherwise set $x_{t+h}^i = x_t^i$. Here $\delta_{x_t^i}(\cdot)$ is a delta function.

We provide a detailed understanding of this inference process as follows. In the second step, λ^i can be interpreted as the intensity with which the probability mass at x_t^i flows to other states $x^i \neq x_t^i$. The probability that x_t^i will change at the current timestep is determined by comparing the threshold $1 - e^{-h\lambda^i}$ with a uniform random variable Z_{change}^i : the larger λ^i is, the more likely a jump will occur. If a change happens, x_{t+h}^i is sampled from all other possible states according to the distribution proportional to $u_t^i(\cdot, x_t^i | x_1^i)$, as defined in Eq. 5. This means the update tends to move x_{t+h}^i towards states that are closer to the model’s prediction x_1^i . In this way, our sampling process enables the model to: (1) continuously refine its predictions along the probability path, and (2) flexibly adjust tokens towards semantically similar alternatives at each timestep. As shown in Fig. 3, this is in contrast to previous mask-based discrete diffusion models [32, 31, 41], where once a token is unmasked, it generally cannot be modified again, even if it contains an error.

4. Experiments

4.1. Implementation Details

In both training stages, we use approximately 13M supervised finetuning data to learn our FUDOKI, including 9M in-house generation data for text-to-image generation and 4M public understanding data, which covers various aspects including OCR [57, 58], doc [59], chart [60], screen [61], math [62, 63], language [64], etc. This is less than Chameleon’s 1.4B data [50] and LWM’s 1B data [65]. We leave the detailed dataset collections in the appendix. For text generation, the sequence length for the response is set to 500, while for image generation, it is set to 576 to match the input size of the image encoder. The text embeddings for calculating the metric distance function $d(\cdot, \cdot)$ are taken from the original embedding layer of Janus-Pro-7B [22] and the image embeddings are obtained from the codebook of LlamaGen [56]. We set $\beta_t = c \left(\frac{t}{1-t}\right)^\alpha$ with $c = 3$ and $\alpha = 0.9$, as suggested in [34]. Besides, following previous studies [41, 40], for the text modality, we pad each sequence with `<eos>` (end-of-sequence) and `<pad>` tokens to the maximum length during training, and compute the loss over model’s answer tokens, including these special tokens. After the sampling process, we only keep the model responses ahead of the first `<eos>` token. The sampling iterations are set as 32 by default, and the resolution of generated images by FUDOKI is 384×384 . The entire training process spanned approximately 43,000 GPU hours.

4.2. Comparison with State-of-the-arts

Visual Generation Performance. We evaluate the generation capabilities of FUDOKI on the widely used GenEval benchmark [71]. Table 1 presents the summarized comparisons, where FUDOKI achieved competitive overall performance (0.77), matching the top score of prior models in the category of both the generation-only and the understanding-and-generation categories. These results underscore our model’s advantages in accurate multi-object understanding and attribute binding, making it promising for complex visual generation tasks that go beyond simple object depiction. This can be attributed to the discrete flow matching framework of FUDOKI, which allows visual information to integrate in both directions for better layout design of generated images.

Multimodal Understanding. We evaluate the understanding capabilities of FUDOKI on sev-

Table 1 | Visual Generation Performance on the GenEval Benchmark. "Und." and "Gen." denotes "Understanding" and "Generation". [†] denotes models that integrate an external pretrained diffusion model.

Type	Paradigm	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	AR	LlamaGen [56]	0.71	0.34	0.21	0.58	0.07	0.04	0.32
		Emu3-Gen [18]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	Diffusion	LDM [12]	0.92	0.29	0.23	0.70	0.02	0.05	0.37
		SDv1.5 [12]	0.97	0.38	0.35	0.76	0.04	0.06	0.43
		PixArt- α [13]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
		SDv2.1 [12]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
		DALL-E 2 [66]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
		SDXL [67]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
		DALL-E 3 [68]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
		SD3-Medium [14]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Und. and Gen.	AR	SEED-X [†] [69]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
		LWM [65]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
		ILLUME [21]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
		TokenFlow-XL [70]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	AR	Chameleon [50]	-	-	-	-	-	-	0.39
		Janus [20]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	AR+Diffusion	Janus-Pro-1B [22]	0.98	0.82	0.51	0.89	0.65	0.56	0.73
		Show-o [52]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	Diffusion	Transfusion [19]	-	-	-	-	-	-	0.63
		UniDisc [44]	0.92	0.47	0.15	0.67	0.13	0.19	0.42
	Discrete Flow	D-DiT [42]	0.97	0.80	0.54	0.76	0.32	0.50	0.65
		FUDOKI (Ours)	0.96	0.85	0.56	0.88	0.68	0.67	0.77
		+Inference Scaling	0.98	0.95	0.73	0.94	0.88	0.78	0.88

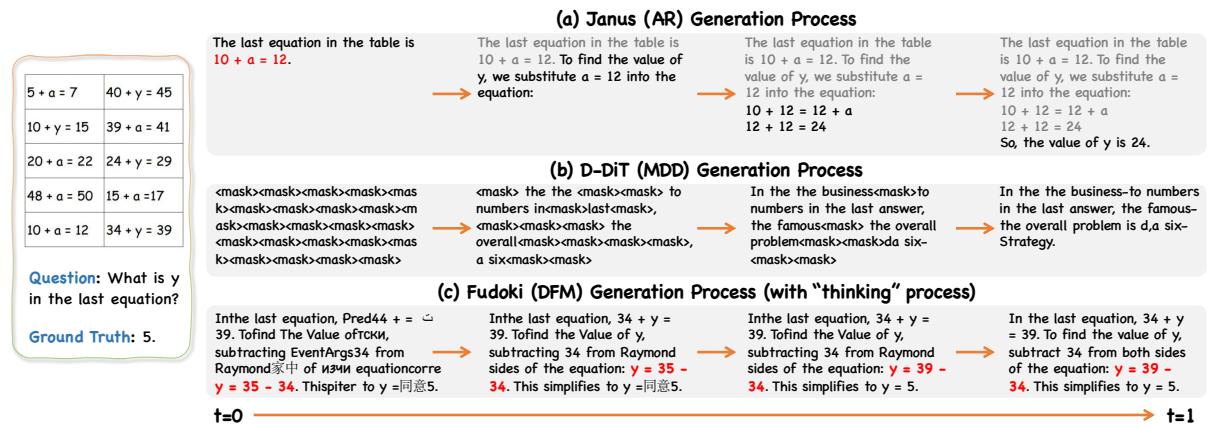


Figure 4 | Generation process of different methods. (a) AR-based Janus can only generate tokens sequentially; if an error is made in the initial step, subsequent outputs will consistently propagate this mistake. (b) D-DiT (mask-based discrete diffusion, MDD) cannot revise tokens once unmasked, making errors irreversible and leading to poor generalization. (c) FUDOKI (discrete flow matching, DFM) allows generated tokens to be revised in subsequent steps, enabling step-by-step reasoning and error correction for more accurate answers.

eral benchmarks, including POPE [82], MME-P [83], SEED [84], MMB [85], GQA [86], MMMU [87], and MM-Vet [88]. Table 2 presents the summarized results ⁵. Notably, our FUDOKI model (1.5B parameters) achieved highly competitive results, which are on par with or surpass several AR-based MLLMs of similar or even larger scale. This demonstrates that FUDOKI delivered

⁵UniDisc [44] is not included in the table due to their inability to conduct visual question answering tasks.

Table 2 | Multimodal Understanding Performance on Various Benchmarks. "Und." and "Gen." denotes "Understanding" and "Generation". \dagger denotes models that integrate an external pretrained diffusion model.

Type	Paradigm	Model	# LLM Params	POPE \uparrow	MME-P \uparrow	MMB \uparrow	SEED \uparrow	GQA \uparrow	MMMU \uparrow	MM-Vet \uparrow
Und. Only	AR	LLaVA-v1.5-Phi-1.5 [52]	1.3B	84.1	1128.0	-	-	56.5	30.7	-
		MobileVLM [72]	1.4B	84.5	1196.2	53.2	-	56.1	-	-
		MobileVLM-V2 [73]	1.4B	84.3	1302.8	57.7	-	59.3	-	-
		MobileVLM [72]	2.7B	84.9	1288.9	59.6	-	59.0	-	-
		MobileVLM-V2 [73]	2.7B	84.7	1440.5	63.2	-	61.1	-	-
		LLaVA-Phi [74]	2.7B	85.0	1335.1	59.8	-	-	-	28.9
		LLaVA [6]	7B	76.3	809.6	38.7	33.5	-	-	25.5
		LLaVA-v1.5 [75]	7B	85.9	1510.7	64.3	58.6	62.0	35.4	31.1
		InstructBLIP [8]	7B	-	-	36.0	53.4	49.2	-	26.2
		Qwen-VL-Chat [76]	7B	-	1487.5	60.6	58.2	57.5	-	-
		IDEFICS-9B [77]	8B	-	-	48.2	-	38.4	-	-
		Emu3-Chat [18]	8B	85.2	1244	58.5	68.2	60.3	31.6	37.2
		InstructBLIP [8]	13B	78.9	1212.8	-	-	49.5	-	25.6
		LaVIT \dagger [78]	7B	-	-	-	-	46.8	-	-
Und. and Gen.	AR	MetaMorph \dagger [79]	8B	-	-	75.2	71.8	-	-	-
		Gemini-Nano-1 [80]	1.8B	-	-	-	-	-	26.3	-
		ILLUME [21]	7B	88.5	1445.3	65.1	72.9	-	38.2	37.0
		TokenFlow-XL [70]	13B	86.8	1545.9	68.9	68.7	62.7	38.7	40.7
		LWM [65]	7B	75.2	-	-	-	44.8	-	9.6
		VILA-U [81]	7B	85.8	1401.8	-	59.0	60.8	-	33.5
		Chameleon [50]	7B	-	-	-	-	-	22.4	8.3
		Janus [20]	1.5B	87.0	1338.0	69.4	63.7	59.1	30.5	34.3
		Janus-Pro-1B [22]	1.5B	86.2	1444.0	75.5	68.3	59.3	36.3	39.8
		Show-o-256 [52]	1.3B	73.8	948.4	-	-	48.7	25.1	-
AR+Diffusion		Show-o-512 [52]	1.3B	80.0	1097.2	-	-	58.0	26.7	-
		D-Dit [42]	2.0B	84.0	1124.7	-	-	59.2	-	-
		FUDOKI (Ours) +Inference Scaling	1.5B	86.1	1485.4	73.9	68.2	57.6	34.3	38.0
			1.5B	-	-	-	-	-	-	55.5

robust multimodal understanding capabilities, which can be attributed to the bidirectional reasoning property of discrete flow matching. Moreover, we provide generation process comparisons for understanding in Fig. 4, which further highlight the advantages of sampling through discrete flow matching for reasoning, *e.g.*, self-correcting the reasoning process for coherency. Our findings highlight the effectiveness and efficiency of FUDOKI, making it a strong alternative to the established AR-based MLLMs.

Inference Scaling. We applied test-time inference scaling techniques [46] to FUDOKI, leveraging a judge model to score multiple candidate outputs and select the highest-scoring responses. The last rows of Table 1 and Table 2 illustrate the impact of inference scaling on visual generation and understanding. For generation, we used the VILA-Judge model [89] to select the top 4 images from 32 candidates per prompt in the GenEval benchmark, resulting in significant performance gains. For understanding, we employed an LLM as the judge to choose the best response from 8 candidates in the challenging MMVet benchmark, where improvements were observed. These results highlight FUDOKI’s potential for further enhancement through reinforcement learning approaches [1, 90].

4.3. Ablation Studies

Training Strategies. 1) *AR Initialization vs Training from Scratch*: As shown in Fig. 5 (left), we compare models initialized with autoregressive (AR) weights [20] against models trained from

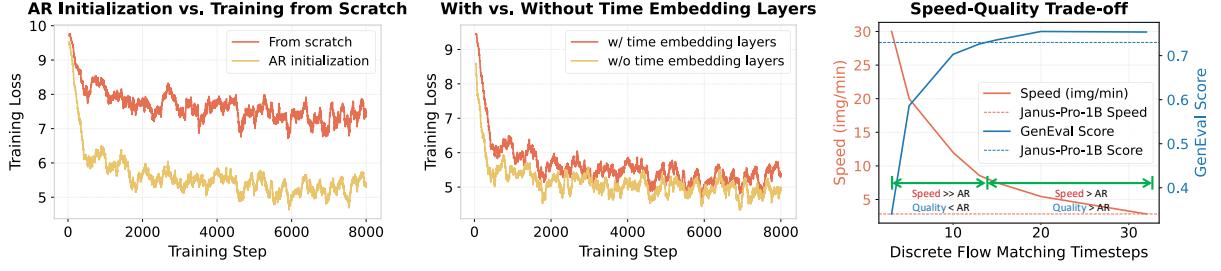


Figure 5 | Comparison of training loss and speed-quality trade-off. **(Left, Middle)** AR initialization and removing time embedding layers both reduce training loss. **(Right)** With fewer timesteps, FUDOKI achieves much higher speed but slightly lower quality than AR; at the optimal timestep, both metrics surpass the AR.

scratch. The results indicate that AR initialization provided a substantial advantage for accelerating model training, leading to consistently lower training loss throughout the optimization process. 2) *Effects of Time-embedding Layers*: We also evaluate the impact of incorporating time embedding layers into the model architecture. The results in Fig. 5 (middle) show that the model without time embedding layers consistently achieves slightly lower training loss than the version with time embeddings. This suggests that our discrete generative model can implicitly infer timesteps from corrupted input, and removing time embeddings reduces model complexity.

Quality-Speed Trade-off. Fig. 5 (right) illustrates the trade-off between speed (in images per minute) and quality (GenEval score) in terms of setting different inference timesteps for visual generation. The red line represents speed, which decreases as the number of timesteps increases, while the blue line represents quality, which improves and stabilizes as timesteps increase. We also draw the dashed horizontal lines indicating the baseline values for the Janus-Pro-1B model (AR), with the red dashed line for speed and the blue dashed line for quality. The plot is divided into two regions, marked by green arrows, illustrating different performance trade-offs: in the first region, speed significantly exceeded the Janus-Pro-1B, but of lower quality; in the second region, the speed of FUDOKI still outperformed the baseline while quality surpassed it. This can be attributed to the significantly fewer NFE (number of function evaluations) and richer bidirectional context modelling of FUDOKI.

5. Conclusion

In this work, we introduced FUDOKI, a multimodal model that uses discrete flow matching to unify visual understanding and generation. Unlike conventional autoregressive and masking-based approaches, FUDOKI leverages discrete flow matching for iterative self-correction, bidirectional reasoning, and flexible generation. Experiments show that FUDOKI performs competitively with leading AR-based MLLMs on both visual understanding and text-to-image generation tasks. These results highlight discrete generative flow models—exemplified by FUDOKI—as a promising direction for advancing multimodal language models and meeting future AGI challenges.

Appendix A. Related Work

Appendix 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 [91, 92], Chameleon [50], AnyGPT [93], 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 [94] 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 [22] decouple visual encoding for understanding and generation to address the granularity mismatch, while VILA-U [81], LWM [65], and LaVIT [51] focus on efficient tokenization, unified visual-text alignment, and scaling to long-context and video scenarios. Illume [21] and Illume+ [48] 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 [95], which enables free-form interleaved multimodal generation; MiniGPT-5 [96], which improves image-text coherence with a two-stage pipeline; NExT-GPT [97], which supports any-to-any modality conversion by connecting AR sequence modeling with modular diffusion decoders; MetaMorph [79], 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 [69], which further enables arbitrary-size and multi-granularity image generation. Recently, BLIP3-o [98] 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 [99] employs a continuous reactified flow for image generation, Show-o [52] 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. Uni-Disc [44] and D-Dit [42] formulate both text and image generation as a discrete diffusion process, starting from masked sequences and enabling joint inpainting, editing, and controllable mul-

timodal 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.

Appendix 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. [38] 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. [37] 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 [100] 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. [35] 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. [33] 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 [101–104], speech and audio generation [105–107], protein design [108–110], and robot control [111]. These successes underscore the broad applicability and effectiveness of flow matching frameworks.

Appendix A.3. Discrete Diffusion Models

Diffusion models have achieved remarkable success in continuous domains such as images and audio [53, 112, 113]. 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 [114, 115]. Representative models in this line include Diffusion-LM [114], DiffuSeq [115], and Plaid [116]. 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 [117] and D3PM [29]. 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 [54], LLaDA [40], and MD4 [31] introduce improvements in noise scheduling, scalability, and training objectives. Methods like MaskGIT [118] and FiLM [119], 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 [30] 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 [120] introduces a reparameterized sampling framework to improve training and sampling efficiency. Furthermore, recent studies [121, 122] model discrete diffusion as a continuous-time Markov chain, advancing theoretical understanding and practical efficiency. Most recently, Discrete Flow Matching (DFM) [33] 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 [123–126, 40, 31]. Recent work [45] 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 [41] further advances diffusion LMs by consistently outperforming previous diffusion models and matching the performance of top autoregressive models of similar size.

Appendix B. More Comparison with State-of-the-arts

Visual Generation Performance on DPG-Bench. We evaluate the visual generation performance of FUDOKI on DPG-Bench [127] (Dense Prompt Graph Benchmark), a comprehensive dataset comprising 1,065 lengthy and densely composed prompts specifically designed to assess the fine-grained semantic alignment capabilities of text-to-image models. As shown in Table 3, FUDOKI demonstrates competitive performance compared to both generation-specialized and unified multimodal models. These results highlight FUDOKI’s strong ability to handle complex, information-rich prompts, establishing it as a robust and versatile solution for multi-aspect visual generation tasks.

Table 3 | Visual Generation Performance on DPG-Bench.

Method	Global	Entity	Attribute	Relation	Other	Overall↑
SDv1.5 [12]	74.63	74.23	75.39	73.49	67.81	63.18
PixArt- α [13]	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next [128]	82.82	88.65	86.44	80.53	81.82	74.63
SDXL [67]	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 [129]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [130]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt- Σ [131]	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen [18]	85.21	86.68	86.84	90.22	83.15	80.60
DALL-E 3 [68]	90.97	89.61	88.39	90.58	89.83	83.50
SD3-Medium [14]	87.90	91.01	88.83	80.70	88.68	84.08
Janus [20]	82.33	87.38	87.70	85.46	86.41	79.68
Janus-Pro-1B [22]	87.58	88.63	88.17	88.98	88.30	82.63
FUDOKI (Ours)	80.55	89.73	88.05	93.66	78.00	83.63

Qualitative Comparisons on Visual Generation. Figure 6 presents qualitative comparisons of visual generation results produced by three models: Janus [20], D-DiT [42], 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) [42], 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.

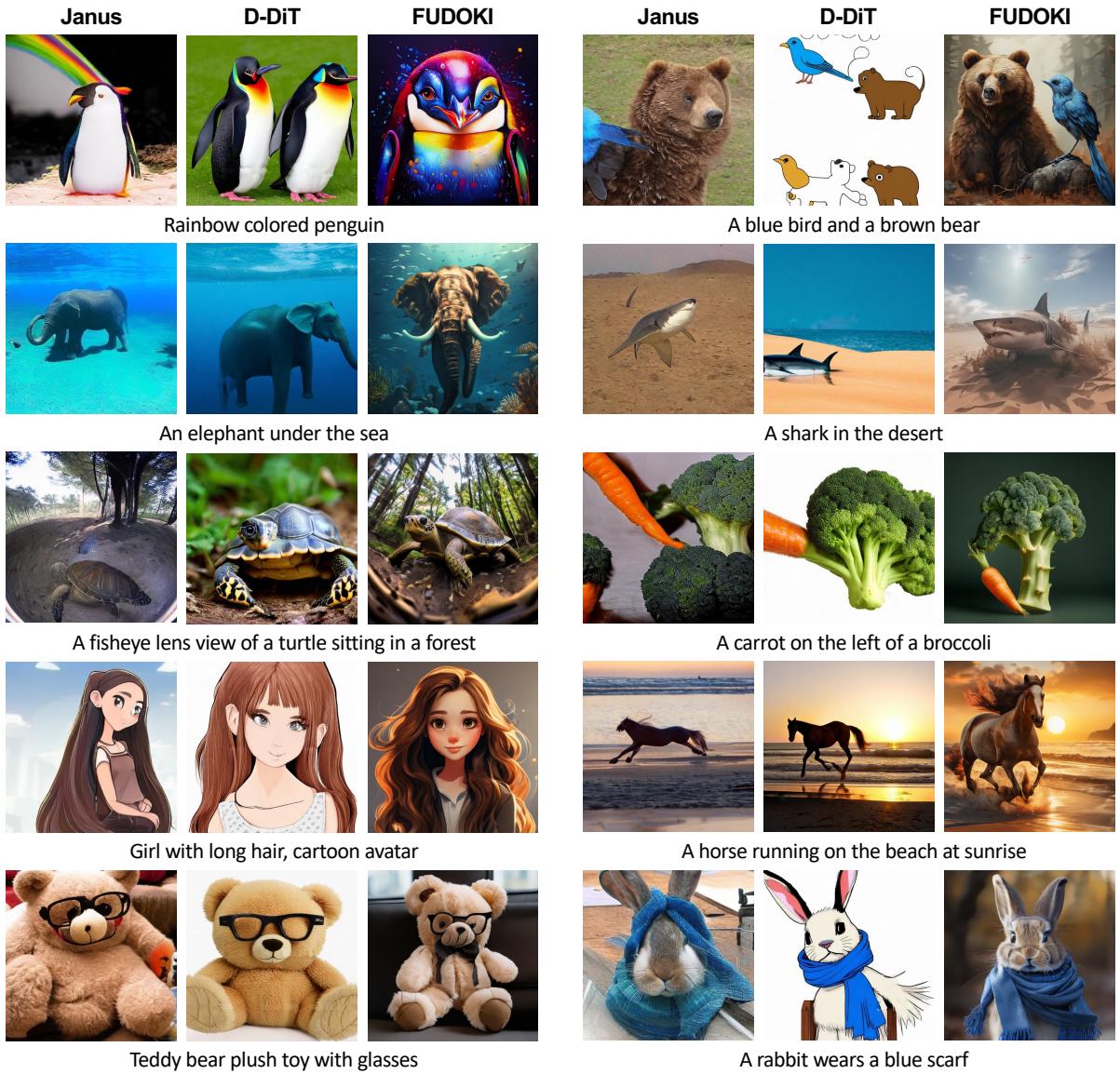


Figure 6 | Qualitative Comparisons on Visual Generation. Comparison among Janus [20], D-DiT [42] and FUDOKI on various text prompts. The results demonstrate that our method (FUDOKI) achieved superior text-image alignment and aesthetics.

Appendix 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,



Figure 7 | Qualitative Comparisons on Visual Understanding. The upper part of the figure shows selected intermediate outputs from the answer generation process of different models—Janus (AR), D-DiT (mask-based discrete diffusion, MDD), and our FUDOKI (discrete flow matching, DFM)—to illustrate their reasoning approaches. Specifically, Janus, the AR-based model, is unable to revise its initial incorrect response (*i.e.*, “Yes, it is summertime ...”), even after generating the correct rationale later (*i.e.*, “The large pumpkins ... suggest that it is autumn”), making its response inconsistent overall. Meanwhile, D-DiT, the mask-based diffusion model, fails to handle this reasoning task, often producing empty outputs (*i.e.*, only </s> tokens). In contrast, our discrete flow matching model, FUDOKI, demonstrates a coherent and accurate reasoning trajectory, producing consistent and correct answers. The lower part of the figure provides additional qualitative examples on visual question answering tasks. FUDOKI consistently delivers more accurate and well-aligned reasoning with the ground truth.

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,

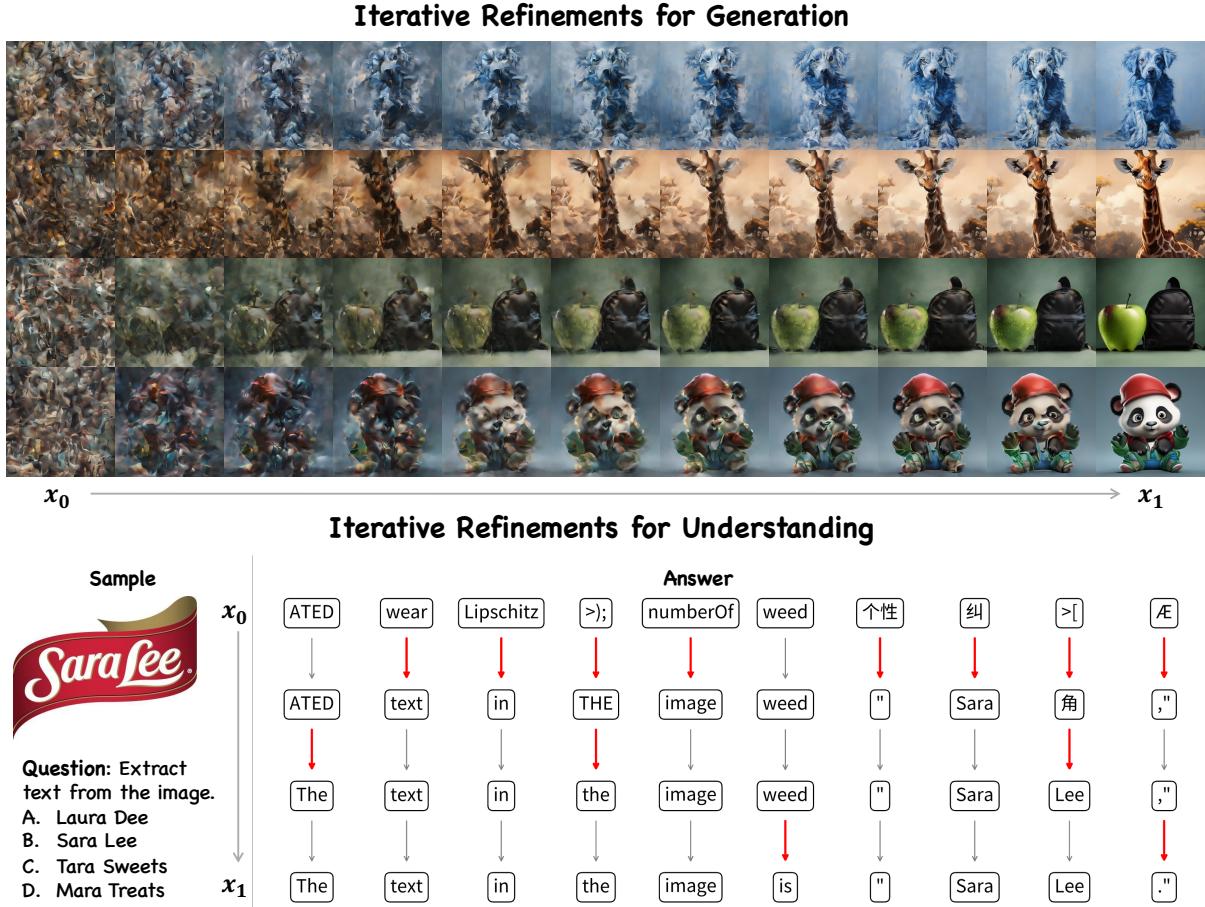


Figure 8 | Visualization of the iterative refinement process enabled by discrete flow matching in FUDOKI, demonstrating denoising process for text-to-image generation and visual understanding tasks.

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, Fig. 9 presents a series of multimodal decision-making scenarios where FUDOKI and GPT-4o/GPT-Image-1 are evaluated on their ability to reason over spatial layouts and produce both textual and visual outputs. Each case involves a frozen lake grid of increasing size (3×3 , 4×4 , and 5×5), with a defined goal and a character’s current position. The task is to select a safe move that avoids hazards (dark blue holes) while progressing toward the treasure. We notice that while GPT-4o provided well-reasoned textual explanations that include safety considerations, goal alignment, and environmental awareness, its visual updates lacked consistency with its textual responses, and even altered the maze structure (in the third row of the figure). In contrast, FUDOKI consistently predicted plausible directions and generated coherent visual updates aligned with the task constraints, showing basic 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.

Appendix 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) [132], VSR (12.8K) [133], ALLaVA-Instruct (680.4K) [134], IconQA (29.9K) [135], LVIS-Instruct4V (10.0K) [136], ShareGPT4V (613.3K) [137], VIQuAE (18.5K) [138], RAVEN (0.3K) [139], Visual7W (14.4K) [140], In-house (70.0K)
- **OCR** (428.0K, 11.5%): LLaVAR (59.3K) [57], SROIE (17.1K) [141], FUNSD (6.8K) [142], OCRVQA (80K) [143], MLHME-38K (30K) [144], Rendered Text (10.0K) [58], IIIT5K (6.0K) [145], HME100K (74.5K) [146], SynthDoG-EN (29.8K) [147], POIE (9.4K) [148], IAM (5.7K) [149], TextCaps (60.5K) [150], COCO-Text V2.0 (28.1K) [151], ChromeWriting (8.8K) [58], ORAND-CAR (2K) [152]
- **Document** (155.8K, 4.2%): DocVQA (122.4K) [59], FUNSD (6.8K) [142], Deepform (9.2K) [153], Kleister CharityAI (15.2K) [154], TAT-DQA (2.2K) [155]
- **Table** (180.2K, 4.9%): TabFact (65.6K) [155], WikiTable (29.5K) [156], TabMWP (38.4K) [157], RoBUT WTQ (38.2K) [158], RoBUT SQA (8.5K) [158]
- **Chart** (362.6K, 9.8%): ChartQA (62.9K) [159], Chart2Text (27.0K) [60], PlotQA (10K) [160], DVQA (200K) [161], Infographic VQA (47.6K) [162], VisText (10.0K) [163], Diagram Image2Text (0.3K) [164], LRV Chart (1.8K) [165]
- **Screen** (24.6K, 0.7%): WebSRC (5.1K) [166], VisualMRC (19.5K) [61]
- **Math/Science** (544.9K, 14.7%): MAVIS (187.3K) [167], G-LLaVA (162.4K) [62], GeoQA+ (72.3K) [63], GeoMVerse (9.3K) [168], Geometry3K (3.0K) [169], MathVision (3.0K) [170], Cambrian Data Engine (50.8K) [171], Textbook QA (21.8K) [172], ScienceQA (19.2K) [173], AI2d (18.8K) [174]
- **Language** (510.2K, 13.7%): MathInstruct (81.5K) [175], Evol-Instruct (142.8K) [176], Math-Plus (95.2K) [177], Magpie Pro (L3 MT) (50.0K) [64], ShareGPT4 (40.7K) [178], Magpie Pro (L3 ST) (50.0K) [64], Magpie Pro (Qwen2 ST) (50.0K) [64]

Appendix 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: $p_0 = p, \quad p_1 = q$

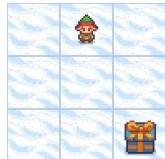
Here, $w_t(x, z) > 0$ is a problem-specific weight controlling the "cost" of mass moving from z to x . As evidenced in [34], when p_t is given and let $w_t(x, z) = 1/p_t(x)$, the kinetic optimal solution can be obtained via $j_t^*(x, z) = [p_t(z)\dot{p}_t(x) - \dot{p}_t(z)p_t(x)]_+ \quad \forall x \neq z$. 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.

Appendix 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.

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.

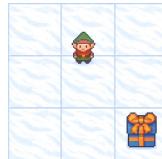
User: Task: Find a safe path through the frozen lake to reach the goal.
 Game Information:
 - You're on a 3x3 frozen lake grid
 - Your goal is to reach the treasure
 - Avoid falling into holes (dark blue areas)
 - Current position: (1, 2)
 - Goal position: (3, 3)
 What direction should you move (Up, Down, Left, Right)?
 Also provide the coordinates you'll reach after this move.
 Then, generate an image of the frozen lake after making this move.



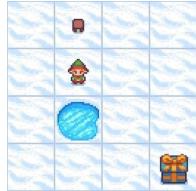
FUDOKI:
 Down
 (2, 2)



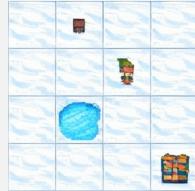
GPT-4o/GPT-Image-1: Based on the current position (1, 2), the best direction to move is Down to (2, 2) to safely progress toward the goal (3, 3). Here is the updated frozen lake grid after the move:
 • The character has moved from (1, 2) to (2, 2).
 • The treasure remains at (3, 3).



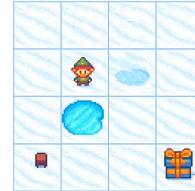
User: Task: Find a safe path through the frozen lake to reach the goal.
 Game Information:
 - You're on a 4x4 frozen lake grid
 - Your goal is to reach the treasure
 - Avoid falling into holes (dark blue areas)
 - Current position: (2, 2)
 - Goal position: (4, 4)
 What direction should you move (Up, Down, Left, Right)?
 Also provide the coordinates you'll reach after this move.
 Then, generate an image of the frozen lake after making this move.



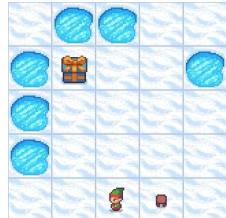
FUDOKI:
 Right
 (2, 3)



GPT-4o/GPT-Image-1: Based on the current position (2, 2), the best direction to move is Right to (2, 3) to safely progress toward the goal (4, 4) while avoiding the hole at (3, 2). Here is the updated frozen lake grid after the move:
 • The character has moved from (2, 2) to (2, 3).
 • The treasure remains at (4, 4).
 • The hole at (3, 2) is still present.



User: Task: Find a safe path through the frozen lake to reach the goal.
 Game Information:
 - You're on a 5x5 frozen lake grid
 - Your goal is to reach the treasure
 - Avoid falling into holes (dark blue areas)
 - Current position: (5, 3)
 - Goal position: (2, 2)
 What direction should you move (Up, Down, Left, Right)?
 Also provide the coordinates you'll reach after this move.
 Then, generate an image of the frozen lake after making this move.



FUDOKI:
 Left
 (5, 2)



GPT-4o/GPT-Image-1: Based on the current position (5, 3), the best direction to move is Up to (4, 3) to safely progress toward the goal (2, 2) while avoiding the holes. Here is the updated frozen lake grid after the move:
 The character has moved from (5, 3) to (4, 3).
 The treasure remains at (2, 2).
 The holes remain in their respective positions.

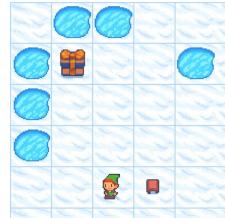


Figure 9 | Comparison of FUDOKI and GPT-4o/GPT-Image-1 on frozen lake maze navigation tasks. GPT-4o/GPT-Image-1 offered well-reasoned textual outputs with safety and goal awareness but generated inconsistent visuals, even altering the maze (e.g., the third row). FUDOKI, by contrast, consistently produced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

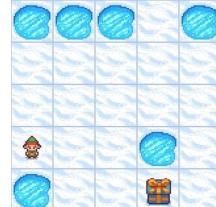
User: Task: Find a safe path through the frozen lake to reach the goal.

Game Information:

- You're on a 5x5 frozen lake grid
- Your goal is to reach the treasure
- Avoid falling into holes (dark blue areas)
- Current position: (4, 1)
- Goal position: (5, 4)

What direction should you move (Up, Down, Left, Right)? Also provide the coordinates you'll reach after this move.

Then, generate an image of the frozen lake after making this move.



FUDOKI: Right (4, 2)



User: Task: Find a safe path through the frozen lake to reach the goal.

Game Information:

- You're on a 5x5 frozen lake grid
- Your goal is to reach the treasure
- Avoid falling into holes (dark blue areas)
- Current position: (4, 2)
- Goal position: (5, 4)

What direction should you move (Up, Down, Left, Right)? Also provide the coordinates you'll reach after this move.

Then, generate an image of the frozen lake after making this move.



FUDOKI: Down (5, 2)



User: Task: Find a safe path through the frozen lake to reach the goal.

Game Information:

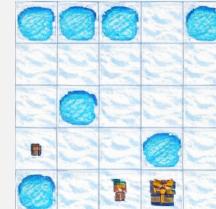
- You're on a 5x5 frozen lake grid
- Your goal is to reach the treasure
- Avoid falling into holes (dark blue areas)
- Current position: (5, 2)
- Goal position: (5, 4)

What direction should you move (Up, Down, Left, Right)? Also provide the coordinates you'll reach after this move.

Then, generate an image of the frozen lake after making this move.



FUDOKI: Right (5, 3)



User: Task: Find a safe path through the frozen lake to reach the goal.

Game Information:

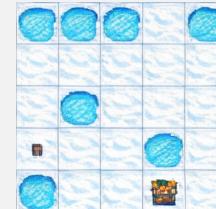
- You're on a 5x5 frozen lake grid
- Your goal is to reach the treasure
- Avoid falling into holes (dark blue areas)
- Current position: (5, 3)
- Goal position: (5, 4)

What direction should you move (Up, Down, Left, Right)? Also provide the coordinates you'll reach after this move.

Then, generate an image of the frozen lake after making this move.



FUDOKI: Right (5, 4)



Success!

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.

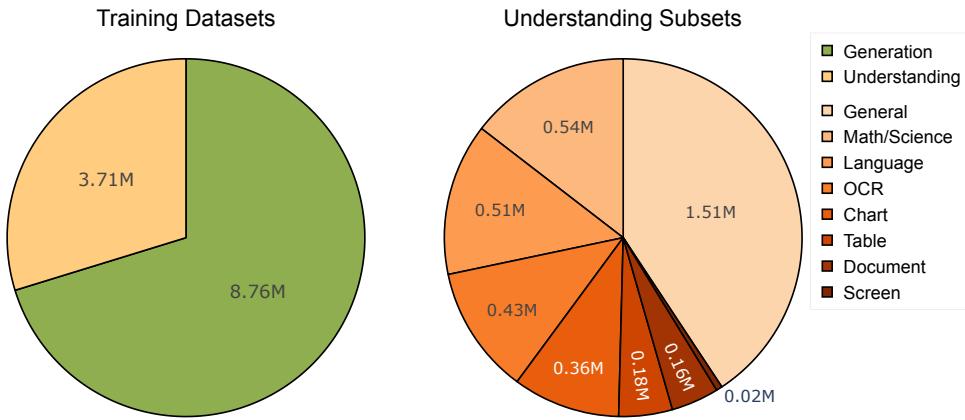


Figure 11 | Training Dataset Distribution. The overall training data consists of 8.76M Generation samples (69%) and 3.86M Understanding samples (31%), as shown on the left. The right chart depicts the composition of the Understanding subset by category.

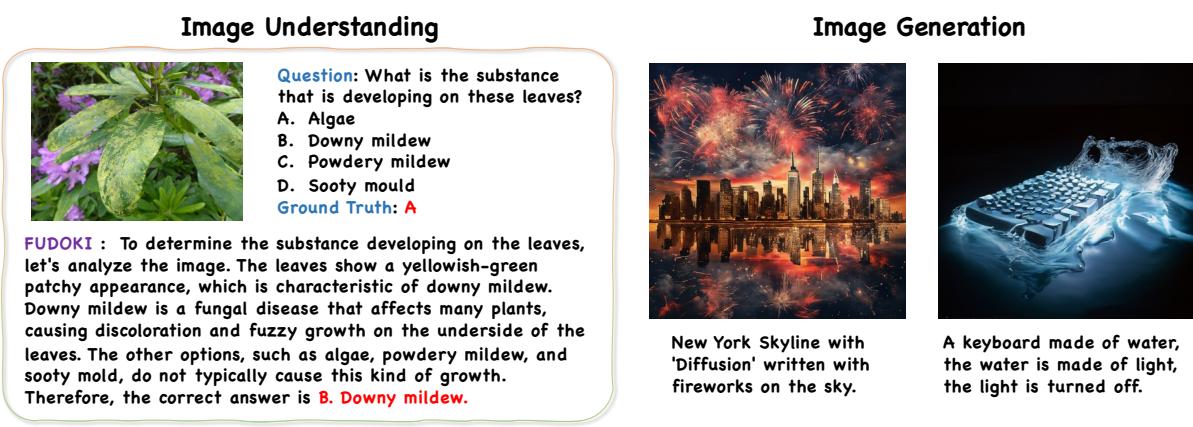


Figure 12 | Examples of failed cases on visual understanding and generation. While FUDOKI demonstrated strong performance, it still struggled with harder tasks—such as generating images from complex prompts involving specific texts, and understanding visuals that require expert-level knowledge.

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, Haojong 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, Zhusuo 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, Yueze Wang, Zhen Li, Qiying 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] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. *ArXiv*, abs/2403.06963, 2024.
- [27] 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.
- [28] 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.

- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] Neta Shaul, Itai Gat, Marton Havasi, Daniel Severo, Anuroop Sriram, Peter Holderith, 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.
- [35] 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.
- [36] Mercury coder, 2025. URL <https://www.inceptionlabs.ai/news>.
- [37] 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.
- [38] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [39] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.
- [40] 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.
- [41] 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>.
- [42] 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.

- [43] 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.
- [44] Alexander Swerdfloor, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [51] 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.
- [52] 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.
- [53] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [54] 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.
- [55] 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.

- [56] 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.
- [57] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.
- [58] Chris Wendler. *wendlerc/renderedtext*, 2023.
- [59] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.
- [60] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model, 2020.
- [61] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, 2021.
- [62] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] 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.
- [67] 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.
- [68] 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.
- [69] 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.
- [70] 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.

- [71] 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.
- [72] 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.
- [73] 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.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] 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.
- [78] 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.
- [79] 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.
- [80] 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.
- [81] 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.
- [82] 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.
- [83] 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.

- [84] 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.
- [85] 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.
- [86] 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.
- [87] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi 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.
- [88] 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.
- [89] 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.
- [90] 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.
- [91] 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.
- [92] 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.
- [93] 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.
- [94] 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.
- [95] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024.
- [96] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.

- [97] 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.
- [98] Juhai 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.
- [99] 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.
- [100] Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR 2023 Conference*, 2023.
- [101] 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.
- [102] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu 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.
- [103] 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.
- [104] 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, Duojun 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, Weiyang Wang, Wenqing Yu, Xinchi 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. Hunyanvideo: A systematic framework for large video generative models, 2025.
- [105] 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.
- [106] 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.

- [107] 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.
- [108] 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.
- [109] 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.
- [110] 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.
- [111] 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.
- [112] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.
- [113] 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.
- [114] 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.
- [115] 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.
- [116] Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715, 2023.
- [117] 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.
- [118] 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.
- [119] Tianxiao Shen, Hao Peng, Ruqi Shen, Yao Fu, Zaid Harchaoui, and Yejin Choi. Film: Fill-in language models for any-order generation. *arXiv preprint arXiv:2310.09930*, 2023.

- [120] Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. In *First Conference on Language Modeling*, 2024.
- [121] 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.
- [122] 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.
- [123] 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.
- [124] Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhengu 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.
- [125] Jiacheng Ye, Zhenyu Wu, Jiahui Gao, Zhiyong Wu, Xin Jiang, Zhengu Li, and Lingpeng Kong. Implicit search via discrete diffusion: A study on chess. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [126] Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhengu Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [127] 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.
- [128] 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.
- [129] 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.
- [130] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi 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.
- [131] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhengu 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.
- [132] Shanghai AI Laboratory. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2023.
- [133] Fangyu Liu, Guy Edward Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.

- [134] 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.
- [135] 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.
- [136] 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.
- [137] 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.
- [138] 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.
- [139] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, 2019.
- [140] 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.
- [141] 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.
- [142] Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*, 2019.
- [143] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.
- [144] Mlhme-38k, 2025. URL <https://ai.100tal.com/icdar>.
- [145] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.
- [146] 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.
- [147] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyo Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

- [148] Jianfeng Kuang, Wei Hua, Dingkang 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.
- [149] 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.
- [150] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020.
- [151] 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.
- [152] Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M. Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S. Oliveira. Proceedings of ieee international conference on frontiers in handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 779–784, 2014. doi: 10.1109/ICFHR.2014.136.
- [153] Deepform, 2025. URL https://wandb.ai/stacey/deepform_v1/reports/DeepForm-Understand-Structured-Documents-at-Scale--Vm1ldzoy0DQ3Njg.
- [154] 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.
- [155] 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.
- [156] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.
- [157] 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.
- [158] 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.
- [159] 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.
- [160] 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.

- [161] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.
- [162] 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.
- [163] Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning, 2023.
- [164] 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.
- [165] 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.
- [166] 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.
- [167] 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.
- [168] 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.
- [169] 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.
- [170] 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.
- [171] 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.
- [172] 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.
- [173] 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.

- [174] 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.
- [175] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.
- [176] Chandeepta 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.
- [177] Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhua Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660, 2024.
- [178] Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Dixin 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.