End-to-End Multi-Modal Diffusion Mamba

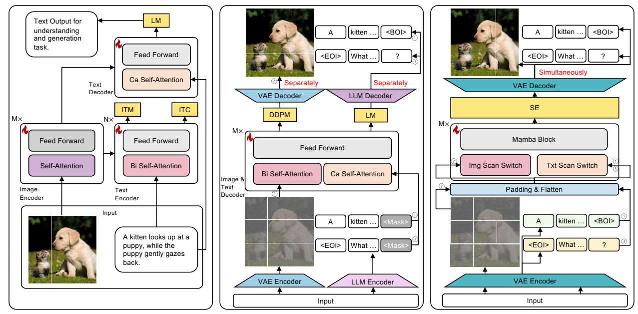
Chunhao Lu[1] Qiang Lu[1✉] Meichen Dong[1,2] Jake Luo[3]
[1]China University of Petroleum-Beijing, [2]Leyard Optoelectronic, [3]University of Wisconsin-Milwaukee
{luchunhao, meichen.dong}@student.cup.edu.cn luqiang@cup.edu.cn jakeluo@uwm.edu

Abstract

Current end-to-end multi-modal models utilize different encoders and decoders to process input and output information. This separation hinders the joint representation learning of various modalities. To unify multi-modal processing, we propose a novel architecture called MDM (Multi-modal Diffusion Mamba). MDM utilizes a Mamba-based multi-step selection diffusion model to progressively generate and refine modality-specific information through a unified variational autoencoder for both encoding and decoding. This innovative approach allows MDM to achieve superior performance when processing high-dimensional data, particularly in generating high-resolution images and extended text sequences simultaneously. Our evaluations in areas such as image generation, image captioning, visual question answering, text comprehension, and reasoning tasks demonstrate that MDM significantly outperforms existing end-to-end models (MonoFormer, LlamaGen, and Chameleon etc.) and competes effectively with SOTA models like GPT-4V, Gemini Pro, and Mistral. Our results validate MDM's effectiveness in unifying multi-modal processes while maintaining computational efficiency, establishing a new direction for end-to-end multi-modal architectures.

1. Introduction

Traditional large-scale multi-modal models [2, 4, 17, 21, 41, 43, 46, 48, 58, 62, 64, 86] typically use multiple encoders and decoders to process multi-modal data. This approach makes learning a unified joint representation of the multi-modal data difficult and can significantly slow inference time (as shown in Fig. 1A). To alleviate these problems, end-to-end models without modal-fusion en(de)coder architecture have been proposed (as shown in Fig. 1B). This approach offers a streamlined, unified processing framework that enhances efficiency and consistency in multi-modal representation learning. Existing end-to-end models follow three primary strategies: (1) Autoregressive models [5, 31, 69, 71] leverage a single Transformer for both text and image generation, but struggle with the inherent sequential dependency of autoregressive decoding. (2) Hybrid image generation models [23, 78] integrate an additional image synthesis module, improving image quality but introducing extra complexity. (3) Mixed autoregressive-diffusion models [89, 90] employ diffusion-based image generation while maintaining an autoregressive framework for text, yet still struggles with unifying multi-modal.



(A) Traditional Multi-Modal Model (B) End-To-End Multi-Modal Model (C) Ours

Figure 1. Comparison of three types of models.

Despite recent advancements, Transformer-based end-to-end models face several critical challenges: (1) their quadratic computational complexity makes them inefficient for generating high-resolution image and long-sequence text. Although various studies have attempted to optimize this computational complexity [1, 3, 12, 27, 29, 56, 59, 66, 74, 75], the challenge remain substantial. (2) their reliance on multi-objective learning introduces conflicting optimization goals, impeding convergence and hindering effective joint representation learning. In contrast, state-space models like Mamba [26, 61] offer a compelling alternative due to their ability to scale linearly with sequence length while effectively capturing long-range dependencies. However, the current multi-modal implementations of Mamba [18, 22, 30, 37, 49, 60, 73, 76, 80, 82] still adopt a multi-objective approach, limiting their capacity for end-to-end joint representation learning.

To effectively process multi-modal data, we propose an end-to-end model called the Multi-Modal Diffusion Mamba (MDM) (as shown in Fig. 1c). MDM first employs patchify [19] and embedding to pre-process multi-modal data. Then, it uses a variational autoencoder (VAE) [42] as a multi-modal encoder, which uniformly maps the multi-

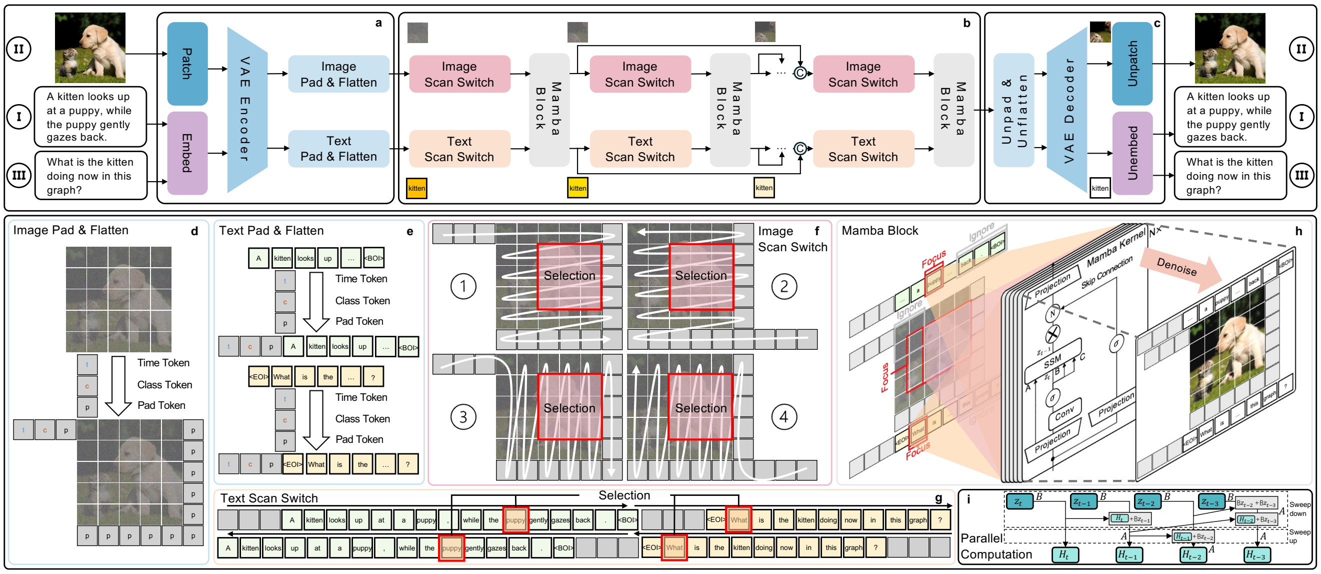


Figure 2. Framework of Multi-Modal Diffusion Mamba. MDM first encodes inputs (caption, VQVAE-processed image, question) using VAE (a), while performing padding (class, diffusion timestep, token completion) and flatten operations (d, e). Next, data reconstruction is progressively completed via diffusion mamba operations (b), modeling images and text temporally through scanning processes (f, g) for efficient information selection (red boxes indicate selection). Selected data undergoes computation (i) guided by (h) within the Mamba-2 framework to update model parameters. Finally, the MDM output passes through the VAE decoder (c) to reconstruct real data.

modal data to a noisy latent space (as illustrated in Fig. 2a). MDM constructs a multi-step selection diffusion model based on the Mamba architecture as a uniform decoder for the rapid generation of multi-modal information.

This decoder generates the target text or image step-by-step based on the diffusion process through the multi-step selection diffusion model (as shown in Fig. 2b). To enhance decoding speed, the decoder employs the Score Entropy Loss [50] as the objective function instead of Markov chain-based [35] methods for updating the network to handle multi-modal data throughout the diffusion process. The decoder comprises two components: an image and text scan switch, and a Mamba-2 block [26]. The text scan switch has two modes for sequence modeling (as shown in Fig. 2f), while the image scan switch has four, based on the settings of DiM [73] (as shown in Fig. 2e). The scan switches enable the model to capture sequential relationships across various temporal directions in the data. The selection state-space structure in Mamba then analyzes these sequential relationships within the current denoising step. This analysis guides the selection of relevant information to focus on and irrelevant information to ignore, effectively directing the model's denoising process at each step.

Since MDM unifies the modality encoder and decoder, the model is capable of generating an image and text simultaneously. For example, as shown in Fig. 2h, when generating an image of a dog alongside its description, the scan switch in the decoder first assesses whether the description contains conditions that necessitate image generation. If such conditions exist, the image scan switch is activated. Consequently, the model directs its selection to the image patches corresponding to the dog during each denoising step. This targeted focus guides the model to effectively denoise relevant pixels while disregarding other areas of the image. A similar selection process is employed for text data. Ultimately, the data, once denoised via the t-step diffusion process, is reconstructed into authentic text (or an image) through the VAE decoder simultaneously. The main contributions of this paper are as follows.

1) We introduce the Multi-Modal Diffusion Mamba (MDM), an end-to-end model that achieves a computational complexity of $\mathcal{O}(MLN^2)$, outperforming previous end-to-end models like MonoFormer [89], which operate at $\mathcal{O}(ML^2N/G)$. This advancement enables the efficient generation of long-sequence text and high-resolution images.

2) We propose a novel multi-step selection diffusion model that combines autoregressive and diffusion-based generative paradigms into a unified learning objective. This method effectively integrates both paradigms within a diffusion process, generating multi-modal data simultaneously.

3) Our experimental results demonstrate MDM's superior performance in image generation on the ImageNet [13] and COCO datasets [40]. Additionally, it excels in various tasks, including image captioning on Flickr30K [84] and COCO [40], VQA on VQAv2 [25], VizWiz [28], and OKVQA [53], as well as text comprehension and reasoning on seven datasets [7, 9, 10, 54, 65, 87]. Furthermore, MDM shows strong results in math-related world knowledge tasks on GSM8k [11], MATH [33], and MMLU [32].

2. Related Works

2.1. Traditional large multi-modal model

Most existing LMMs are built by integrating architectures from multiple modalities [2, 4, 41, 46, 62, 86]. SOTA image and video generation models employ pre-trained text encoders to represent input prompts in latent space, which then condition a diffusion model for generating videos and images [64]. Many researchers have adopted this approach, fusing feature representations from multiple pre-trained encoders to enhance model performance across different modalities [21, 58]. This pattern is also prevalent in visual language models, where pre-trained language models are typically augmented with linear projection layers from other pre-trained en/decoders for training in the text space. Examples include Flamingo [2] and LLaVA [48] for visual understanding, GILL [43] for visual generation, and DreamLLM [17] for both understanding and generation.

2.2. End-to-End multi-modal model

End-to-end models have emerged recently to facilitate joint representation learning while improving training and inference efficiency. It can be categorized into three main types:
1) **The autoregressive model** [5, 31, 69, 71] utilizes one Transformer with an autoregressive approach to generate images and text. For instance, the Fuyu model [5] processes image patches directly as input to achieve visual comprehension. Models like Chameleon [71], Mars [31], and LlamaGen [69] convert images into discrete sequence tokens, then concatenate them with text.
2) **The hybrid image generation model** [23, 78] addresses the limitations of autoregressive approaches in image generation. While maintaining an autoregressive structure for text generation, the models enhance image quality by incorporating an image-generation network. For example, Seed-x model [23] focuses on enhancing specific aspects of image generation, while Next-GPT [78] aims to expand multi-modal capabilities within an end-to-end framework.
3) **The mixed autoregressive-diffusion model** [89, 90] combines the strengths of previous approaches. It performs text autoregressive generation and image diffusion restoration simultaneously. Models like MonoFormer [89] and Transfusion [90] achieve this by incorporating causal self-attention [81] for text tokens and bidirectional self-attention [14] for image patches, enabling high-quality multi-modal understanding and generation.

2.3. Mamba in multi-modal model

Mamba has emerged as a powerful alternative to Transformer for multi-modal data alignment [18, 49, 76, 77, 82]. Recent works showcase Mamba's capabilities across different multi-modal applications. VL-Mamba [60] combines a pre-trained Mamba model for language understanding with a connector module to align visual patches and language tokens. However, these models lack end-to-end training capabilities and struggle to learn unified joint representations. MDM provides a truly end-to-end architecture, enabling rapid generation of high-quality, long sequences.

3. Multi-step Selection Diffusion Model

The multi-step selection diffusion model enables rapid generation of multi-modal information through two key processes: diffusion & denoising and selection. During the diffusion & denoising, the model employs a unified Score Entropy Loss [50](SE) to gradually reconstruct target data from noise through a series of denoising steps (as illustrated in Fig. 2b). The selection process enables the model to capture sequential relationships across different temporal dimensions in the latent space, determining which information should be focused on or ignored during each diffusion denoising step (as shown in Fig. 2h).

3.1. Diffusion & Denoising

The diffusion & denoising process comprises two main components: diffusion and denoising. The diffusion component can be expressed by the following equation:

$$z^g_{n,t} = \sqrt{\bar{\alpha}^g_t}\, z^g_{n,0} + \sqrt{1-\bar{\alpha}^g_t}\, \epsilon^g_{n,t}, \tag{1}$$

where g denotes either image patch or text embedding, and $z^g_{n,0}$ represents the latent space vector of the n-th image patch or text embedding, obtained through VAE sampling [42]. $z^g_{n,t}$ is derived from $z^g_{n,0}$ after t steps of noise addition; $\epsilon^g_{n,t} \sim \mathcal{N}(0, I)$ represents the added noise; $\bar{\alpha}^g_t = \prod_{k=1}^{t} \alpha^g_k$, $\alpha^g_k = 1 - \beta^g_k$, and $\{\beta^g_k \in (0,1)\}^T_{k=1}$ are Gaussian distribution hyperparameters controlling the forward diffusion noise. Following the diffusion Markov principle [35], t-step forward diffusion process can be characterized by conditional probabilities as follows:

$$p(z^g_{n,t} | z^g_{n,0}) = \mathcal{N}(z^g_{n,t}; \sqrt{\bar{\alpha}^g_t}\, z^g_{n,0}, (1-\bar{\alpha}^g_t) I), \tag{2}$$

which means that given $z^g_{n,0}$, $z^g_{n,t}$ follows a Gaussian distribution with $\sqrt{\bar{\alpha}^g_t}\, z^g_{n,0}$ as mean and $(1-\bar{\alpha}^g_t)I$ as variance.

In the classic diffusion denoising component [35], the model needs to learn the posterior $p(z^g_{n,t-1} | z^g_{n,t})$ to gradually reconstruct the data. Since $p(z^g_{n,t} | z^g_{n,0})$ follows a Gaussian distribution, we can assume that the approximate distribution of the denoising process is:

$$p_\theta(z^g_{n,t-1} | z^g_{n,t}) = \mathcal{N}(z^g_{n,t-1}; \mu_\theta(z^g_{n,t}), (\sigma^g_{\theta,n})^2). \tag{3}$$

where $\mu_\theta(z^g_{n,t})$ and $\sigma^g_{\theta,n}$ represent the model predicted noise mean and variance at the t-th denoising step.

This method achieves the gradual recovery of data by optimizing the conditional probability of each time step by maximum likelihood. However, Markov chain-based [35] methods limit computational efficiency in high-dimensional spaces and are difficult to extend to discrete data.

To further optimize the denoising process, this paper uses SE [50] as the optimization target. It is a generalized score matching objective that aims to directly learn the probability density ratio between discrete states. The SE can not only stabilize the diffusion denoising process but also improve the sampling quality through the global information of data distribution. In general form, for any state pair $(z^g_{n,t}, z^g_{n,0})$, define the model's score ratio $s_\theta(z^g_{n,t})$, which represents the relative probability of transferring from $z^g_{n,t}$ to $z^g_{n,0}$. SE is defined as:

$$se = \sum_{y \in z^g_{n,0:t-1}} \omega^g_{z^g_{n,t}} \left(s_\theta(z^g_{n,t}) - \frac{p_{data}(y)}{p_{data}(z^g_{n,t})} \log s_\theta(z^g_{n,t}) + K\left(\frac{p_{data}(y)}{p_{data}(z^g_{n,t})}\right) \right), \tag{4}$$

where $\omega^g_{z^g_{n,t}}$ is the weight of the loss term, which is used to balance the loss of different states. $K(a) = a(\log a - 1)$ is a normalization term that ensures the loss is non-negative. $\frac{p_{data}(y)}{p_{data}(z^g_{n,t})}$ represents the actual score ratio. $p_{data}(y)$ and $p_{data}(z^g_{n,t})$ are the actual data distributions of the former noisy state and the current noisy state. The actual score ratio calculation relationship is shown in Theorem 1.

Theorem 1. *According to Bayes' theorem and the Gaussian distribution density formula, the following calculation relationship of* $\frac{p_{data}(y)}{p_{data}(z^g_{n,t})}$ *is obtained:*

$$\frac{p_{data}(y)}{p_{data}(z^g_{n,t})} = \exp\left(\frac{\|z^g_{n,t}\|^2}{2} - \frac{\|z^g_{n,t} - \sqrt{\bar{\alpha}^g_t} z^g_{n,0}\|^2}{2(1-\bar{\alpha}^g_t)}\right). \tag{5}$$

The proof is provided in Appendix A.

Based on the SE [50], the model predicted score ratio indicates how the model adjusts the probability of the current state to tend to the original data distribution during the denoising process. The definition is as follows:

$$s_\theta(z^g_{n,t}) = \frac{p_\theta(z^g_{n,0})}{p_\theta(z^g_{n,t})}, \tag{6}$$

where the denominator represents the probability of the current noise state and the numerator represents the original state probability estimated by the model. According to Theorem 2, the model uses $softmax$ for normalization ensuring numerical stability and enabling gradient optimization when predicting the score ratio.

Theorem 2. *Given the denoising process modelled by a score-based probability ratio function* $s_\theta(z^g_{n,t})$*, defined as Eq.* (6)*, this paper defines a learnable approximation using a parameterized score function* f_θ*, such that the probability ratio can be estimated as:*

$$s_\theta(z^g_{n,t}) = \frac{\exp\left(f_\theta(z^g_{n,t}, z^g_{n,0})\right)}{\sum_{y \in z^g_{n,0:t-1}} \exp\left(f_\theta(z^g_{n,t}, y)\right)}, \tag{7}$$

The proof is provided in Appendix A.

3.2. Selection

The selection process comprises two key steps: scan switch and selection. The scan switch mechanism captures temporal relationships between adjacent image patches (or text embeddings) by generating latent space representations with k different sequential relationships, such as four image patch sequences and two text embedding sequences illustrated in Fig. 2fg. The mechanism creates k temporal sequences $S = \{\langle z^g_{1,t}, z^g_{2,t}, \dots, z^g_{i,t}\rangle\}_k$.

The selection step then analyzes these different sequential relationships at the current denoising step t to determine which information should be focused on or ignored, thereby guiding the model's denoising direction in each diffusion step. The selection step chooses j items $z^g_{n,t}$ from each sequence in S according to the following Theorem 3. So, the selection step obtain k selection sequences with different lengths, i.e., $S' = \{\langle z^g_{j_1,t}, z^g_{j_2,t}, \dots, z^g_{j,t}\rangle\}_k$ and $S' \in S$.

Theorem 3. *To achieve the optimal score entropy [50] which is demonstrated on Eq.* (4)*, the selection step choose* j *items where each* $z^g_{n,t}$ *satisfies* $se = 0$*, i.e.,*

$$s_\theta(z^g_{n,t}) \approx \frac{p_{data}(y)}{p_{data}(z^g_{n,t})} \tag{8}$$

The proof is provided in Appendix A.

4. Architecture

The neural network architecture consists of two primary components: a VAE noisy latent encoder [42] and a multi-step selection diffusion decoder, as illustrated in Fig. 2ab. The encoder first processes image data X_{img} through patchify [19] operations and processes text data X_{txt} through tokenization based on SentencePiece with Unigram BPE [45] and embedding operations, then uniformly maps them to the latent space before applying forward noise.

The decoder, based on the multi-step selection diffusion model, leverages Mamba to achieve unified learning objectives while enhancing computational efficiency for processing long sequence data. It employs the SE [50] as the unified objective for both image and text modalities during the diffusion process. During selection, the model captures sequential relationships across different temporal dimensions using various scan switches. These relationships are then efficiently processed through the selection state-space structure in the Mamba Block determining which information to focus on or ignore according to Eq. (8), thereby guiding subsequent diffusion denoising steps (as shown in Fig. 2h). Finally, the reconstructed image patches and text embeddings are transformed back into their original data formats through a VAE noisy latent decoder [42].

4.1. The noisy latent encoder

The noisy latent encoder first processes input image X_{img} through patchify and processes text X_{txt} through tokenization and embedding operations to obtain the patch sequence

$G(X_{img}/X_{txt}) = \langle g_1, g_2, \ldots, g_i \rangle$, where g_n represents the n-th image patch or text embedding, respectively. The encoder VAE [42] generates Gaussian distribution parameters (mean μ and variance σ) for these patches, with a similar process applied to text embeddings, i.e., $VAE(G) = (\mu, \sigma)$. For each image patch or text embedding g_n, its noise z_n is a sample s_n from the distribution $\mathcal{N}(\mu, \sigma)$ with the addition noise $\epsilon_n \sim \mathcal{N}(0, 1)$, i.e, $z_n = s_n + \epsilon_n$. Finally, the image X_{img} and text X_{txt} are transformed into the noise sequence $\langle z_1, \cdots, z_i \rangle$ through the above process.

Moreover, three types of learnable padding tokens, time, category, and pad, are inserted into these noise sequences, as illustrated in Fig. 2de. The time token encodes the current diffusion step, the class token is used to learn the data category, and the pad token represents the start or end position for splitting these noise sequences.

4.2. The multi-step selection diffusion decoder

The decoder aims at progressively recovering the image X_{img} or text X_{txt} from noise sequences through two main modules: 1) the multi-step selection diffusion Mamba and 2) the VAE noisy latent decoder. 1) The Mamba is used to recover the patch sequence $\langle g_1, \cdots, g_i \rangle$ from the noise sequence $\langle z_1, \cdots, z_i \rangle$. 2) The VAE noisy latent decoder assembles patches and generates the image $\hat{X}_{img}$ or text $\hat{X}_{txt}$.

4.2.1. Multi-step selection diffusion Mamba

The module leverages two components, image/text scan switch and Mamba Block, to implement each denoising step in the multi-step selection diffusion model (Sec. 3).

The image/text scan switch component establishes sequences with different directions to capture different temporal relationships between patches. Following Dim [73], we implement four distinct scan switches for images (as shown in Fig. 2f) and two for text (as shown in Fig. 2g).

The Mamba block is used to select patches from these different scan switch sequences and denoise the input noise $z^g_{n,t}$. The block adopts the state space architecture from Mamba-2 [26]. According to Sec. 3.2, it is s_θ, where $\theta = \{H^g_{n,t}, A, B, C, D, \Delta\}$ represent the state space in the block. The block comprises six key components: 1) linear input and output projection layers, 2) convolution kernel layer, 3) nonlinear activation layer, 4) state space model (SSM), 5) skip connection layer, and 6) normalization layer.

1) The linear input projection layer reduces the dimensionality of the latent space noise vector while simultaneously applying initial state matrices A, B, C to the linear projection of input data $z^g_{n,t}$. Additionally, the linear output projection layer represents the denoising step, which transforms the selection noise $z^g_{n,t}$ into $z^g_{n,t-\Delta t}$ and outputs it to the next Mamba block according to the following equation.

$$z^g_{n,t-\Delta t} = z^g_{n,t} - \frac{\Delta t}{2}[f_\theta(z^g_{n,t}, t) + f_\theta(z^g_{n,t-\Delta t}, t-\Delta t)] \quad (9)$$

where the equation adopts the second-order numerical method of DPM-Solver [51] to improve sampling accuracy. Details are provided in Appendix B.

2) The convolution kernel layer implements parallel scan switches, routing the initial linear projection of the input and the state matrix's linear projection through the SSM, as shown in Fig. 2i. The sweep down and sweep up [26] enable parallel computation between Eqs. (10) to (13).

3) The nonlinear layer enhances model generalization.

4) The SSM lets the Mamba block s_θ approximate the actual score ratio based on Theorem 3. To implement the target, SSM updates the state space θ by the following equations (based on Theorem 3 and details in Appendix A).

$$H^g_{n,t} = \bar{A}H^g_{n,t-1} + \bar{B}z^g_{n,t} \quad (10)$$

$$z^g_{n-1,t} = CH^g_{n,t} + Dz^g_{n,t} \quad (11)$$

$$\bar{A} = \exp(\Delta A) \quad (12)$$

$$\bar{B} = (\Delta A)^{-1} \cdot (\exp(\Delta A) - I) \cdot \Delta B \quad (13)$$

where $H^g_{n,t}$ represents the hidden state representation, A and B control the evolution of hidden states and latent space noise vector inputs, respectively, C governs the hidden state representation of the target output and D manages the nonlinear skip connection for latent space noise vector inputs. Δ denotes the learnable time parameter.

5) The skip connection layer facilitates input feature reuse and mitigates model degradation.

6) The Normalization layer ensures training stability.

According to Eq. (8) in Theorem 3 and Eq. (4), the goal of training the Mamba block is:

$$L_{se} = \mathbb{E}_{z^g_{n,0} \sim p_0, z^g_n \sim p(\cdot|z^g_{n,0})} se = 0 \quad (14)$$

4.2.2. The noisy latent decoder

After applying the diffusion-based denoising process, the recovered latent variable $z^g_{n,0}$ is passed to the VAE decoder [42] as illustrated in Fig. 2c. For image reconstruction, the decoder applies an ℓ_2 loss:

$$L^{img}_{rec} = \mathbb{E}_{z^g_{n,0} \sim q_\phi(z|X)} \|X_{img} - \hat{X}_{img}\|^2. \quad (15)$$

where $q_\phi(z|X)$ represents the posterior distribution of the VAE encoder.

For text, the decoder minimizes the cross-entropy loss:

$$L^{txt}_{rec} = -\mathbb{E}_{z^g_{n,0} \sim q_\phi(z|X)} \sum_t p(X^{(t)}_{txt}|z^g_{n,0}) \log p_\psi(\hat{X}^{(t)}_{txt}|z^g_{n,0}). \quad (16)$$

where $p(X^{(t)}_{txt}|z^g_{n,0})$ represents the probability distribution of real text data under the condition of latent variable $z^g_{n,0}$.

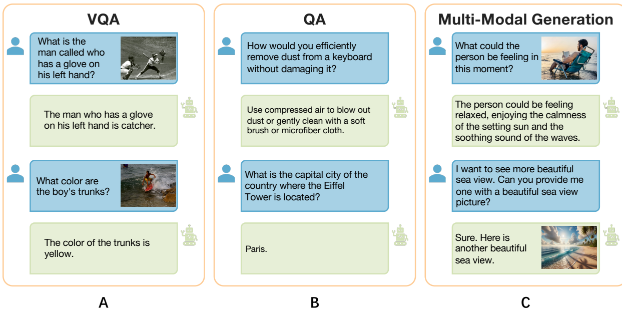


Figure 3. VQA, QA and Multi-Modal generation test from MDM. The results of VQA are part of VQAv2 [25]. The QA results are part of PIQA [7] and MMLU [32]. The Multi-Modal generation results are tested with ground-truth data.

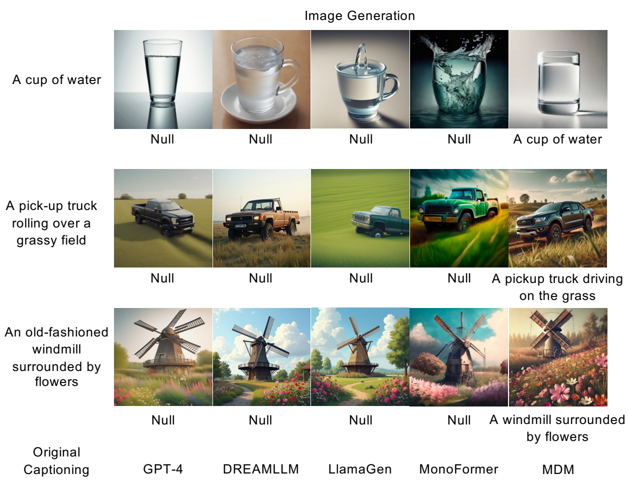


Figure 4. Comparison between each model on generating captioning and image results on COCO dataset. Unlike other models, MDM generates both image and caption data simultaneously.

And $p_\psi(\hat{X}_{txt}^{(t)}|z_{n,0}^g)$ represents the probability distribution of the text token generated by the VAE decoder under the condition of the latent variable $z_{n,0}^g$.

Besides, a KL divergence regularizes the latent space:

$$L_{KL} = D_{KL}\left(q_\phi(z|X)\|p(z)\right). \tag{17}$$

where $p(z)$ represents the prior distribution of the latent variable by VAE, which is assumed to be a standard Gaussian distribution $\mathcal{N}(0, I)$ to regularize the latent variable space and enable it to have smooth generation capabilities.

The final optimization objective integrates VAE reconstruction, KL divergence and SE:

$$L_{total} = L_{rec}^{img} + L_{rec}^{txt} + \beta L_{KL} + \lambda L_{se}. \tag{18}$$

5. Experiments

5.1. Experimental Setup

Model configuration. Our model applies a VAE [42] as the noisy latent encoder and decoder. Moreover, it integrates the DiM selection state space [73] in each Mamba block as the diffusion decoder. The resulting model contains 7 billion parameters, with 49 Mamba blocks in the multi-step selection diffusion decoder, each having a dimension of 2048 (Details of parameter settings listed in Appendix C).

Before the training MDM process, we trained a tokenization model based on SentencePiece (Unigram BPE) [45]. The tokenization model can help the model construct a stable text latent variable representation, thereby optimizing the forward diffusion and reverse denoising process. See Appendix D for detailed experimental settings.

In the training process, we import the DDPM scheduler [35] and DPM-Solver [51] to improve the sampling efficiency in the diffusion model. We then use the AdamW optimizer without weight decay, maintaining a constant learning rate of 0.0001. Meanwhile, we keep an EMA of the model weights with a coefficient of 0.9999.

Baseline and dataset. Our evaluation encompasses four tasks: image generation with classifier-free guidance [34] (CFG), text-to-image, image-to-text, and text-to-text generation. For the baseline model training, we train MDM on ImageNet [13], JourneyDB [68] and UltraChat [16].

For the image generation and the text-to-image task at 256×256 resolution, we compare the MDM baseline model against established baselines across three categories: diffusion models (Imagen [64], ADM [15], CDM [36], LDM [63], DiT-XL/2 [57], SDXL [58], and SD-3 [21]), autoregressive models (VQGAN [20] and ViT-VQGAN [85]), and end-to-end multi-modal models (NExT-GPT [78], Chameleon [71], LlamaGen [69], Transfusion [90], MonoFormer [89], Dual-DiT [47], JanusFlow [52] and Show-O [79]). For the image generation task, we evaluate performance on ImageNet [13] using four metrics: Frechet Inception Distance (FID), Inception Score (IS), and Precision/Recall. For the text-to-image task, we evaluate performance on COCO [40] using FID and Gen Eval [24].

For the image-to-text task (image captioning and vision question answering, VQA) and text-to-text task, we employ MDM baseline model and MDM instruction model by visual instruction tuning [48] on multiple datasets: COCO [40], GQA [38], OCR-VQA [55], TextVQA [67], and VisualGenome [44]. We evaluate the model against two groups of baselines: traditional models and end-to-end multi-modal models. Performance evaluation of image captioning is conducted on Flickr 30K [84] and COCO [40] datasets using the Consensus-based Image Description Evaluation (CIDEr) metric. And performance evaluation of VQA is conducted on VQAv2 [25], VizWiz [28], and OKVQA [53] using answer accuracy rate as the evaluation metric.

For the text-to-text task, we evaluate the model on text comprehension and reasoning tasks using HellaSwag [87], OpenBookQA [54], Wino-Grande [65], ARCEasy, ARC-

Model	Arc	Params	Image Generation with CFG				Text-to-Image Generation	
			FID ↓	IS ↑	Pre ↑	Re ↑	FID ↓	Gen Eval ↑
Imagen [64]	Diff	7.3B	-	-	-	-	7.27	-
ADM [15]	Diff	554M	10.94	101.0	0.69	0.63	-	-
CDM [36]	Diff	-	4.88	158.7	-	-	-	-
LDM [63]	Diff	400M	3.60	147.6	0.87	0.68	-	0.43
DiT-XL/2 [57]	Diff	675M	2.27	278.2	0.83	0.57	-	-
SDXL [58]	Diff	3.4B	-	-	-	-	4.40	0.55
SD-3 [21]	Diff	12.7B	-	-	-	-	-	0.68
VQGAN [20]	AR	227M	18.65	80.4	0.78	0.26	-	-
ViT-VQGAN [85]	AR	1.7B	4.17	175.1	-	-	-	-
NExT-GPT [78]	AR	7B	-	-	-	-	10.07	-
Chameleon [71]	AR	7B	-	-	-	-	26.74	0.39
LlamaGen [69]	AR	3.1B	2.81	311.5	0.84	0.54	4.19	-
Transfusion [90]	AR+Diff	7.3B	-	-	-	-	6.78	0.63
MonoFormer [89]	AR+Diff	1.1B	2.57	272.6	0.84	0.56	-	-
Dual-DiT [47]	Diff	2B	-	-	-	-	9.40	0.65
JanusFlow [52]	AR+Diff	1.3B	-	-	-	-	-	0.70
Show-O [79]	AR+Diff	1.3B	-	-	-	-	9.24	0.68
MDM	Diff	7B	2.49	281.4	0.86	0.59	5.91	0.68

Table 1. Performance on ImageNet and COCO 256×256. FID, IS, Pre, and Re stands for Frechet Inception Distance, Inception Score, Precision, and Recall, respectively.

Model	IC		VQA			Text Comprehension and Reasoning							Math and World		
	Flickr	COCO	VQAv2	VizWiz	OK	HS	OBQA	WG	ARCE	ARCC	BoolQ	PIQA	GSM8k	MATH	MMLU
Llama-2 [74] (7B)	-	-	-	-	-	77.2	58.6	78.5	75.2	45.9	77.4	78.8	14.6	2.5	45.3
Mistral [39] (7B)	-	-	-	-	-	81.3	-	75.3	80.0	55.5	84.7	83.0	52.1	13.1	60.1
Flamingo [2] (80B)	75.1	113.8	67.6	-	-	-	-	-	-	-	-	-	-	-	-
Gemini Pro [72]	82.2	99.8	71.2	-	-	84.7	-	-	-	-	-	-	86.5	32.6	71.8
GPT4V [8]	55.3	78.5	77.2	-	-	95.3	-	-	-	-	-	-	92.0	52.9	86.4
InstructBLIP [48] (7B)	82.4	102.2	-	33.4	33.9	-	-	-	-	-	-	-	-	-	-
mPLUG-Owl [83] (7B)	80.3	119.3	-	39.0	-	-	-	-	-	-	-	-	-	-	-
TinyLlama [88] (1.1B)	-	-	-	-	-	59.2	36.0	59.1	55.3	30.1	57.8	73.3	-	-	-
Pythia [6] (12B)	-	-	-	-	-	52.0	33.2	57.4	54.0	28.5	63.3	70.9	-	-	-
DREAMLLM [17](7B)	-	115.4	56.6	45.8	44.3	-	-	-	-	-	-	-	-	-	-
Emu [70](7B)	-	117.7	40.0	35.4	34.7	-	-	-	-	-	-	-	-	-	-
Chameleon [71](34B)	74.7	120.2	66.0	-	-	74.2	51.0	70.4	76.1	46.5	81.4	79.6	41.6	11.5	52.1
NExT-GPT [78](7B)	84.5	124.9	66.7	48.4	52.1	-	-	-	-	-	-	-	-	-	-
Transfusion [90](7B)	-	33.7	-	-	-	-	-	-	-	-	-	-	-	-	-
MonoFormer [89](1.1B)	-	-	-	-	-	50.6	37.2	56.9	48.2	31.5	62.3	71.2	-	-	-
Dual-DiT [47](2B)	-	56.2	60.1	29.9	25.3	-	-	-	-	-	-	-	-	-	-
JanusFlow [52](1.3B)	-	-	79.8	-	-	-	-	-	-	-	-	-	-	-	-
Show-O [79](1.3B)	67.6	-	74.7	-	-	-	-	-	-	-	-	-	-	-	-
MDM (7B)	62.4	109.6	60.3	39.8	47.1	70.6	41.5	68.8	55.1	46.2	65.7	79.9	40.5	12.1	54.4
InstructMDM (7B)	75.2	122.1	66.7	46.3	51.6	74.8	48.3	74.9	65.4	47.1	71.5	83.7	46.0	13.1	59.2

Table 2. Performance on image-to-text and text-to-text tasks. The evaluation of image captioning (IC) and VQA is CIDEr and answer accuracy % (Flickr is evaluated on 30K and OK represents OKVQA).

Challenge [10], BoolQ [9], and PIQA [7]. We also evaluate the model on math and world knowledge tasks using GSM8K [11], MATH [33], and MMLU [32]. The evaluation metrics for all the tasks are accuracy rates.

5.2. Experimental Results

Image Generation. In the image generation task on ImageNet, MDM achieves top-three rankings across all evaluation metrics: second in FID, IS, and Precision, and third in Recall when compared against one-modal diffusion models and end-to-end multi-modal models (see Tab. 1). MDM demonstrates superior overall performance, notably surpassing other end-to-end multi-modal models in three of the four metrics. In the text-to-image task, we tested the model on the COCO dataset to generate both image and

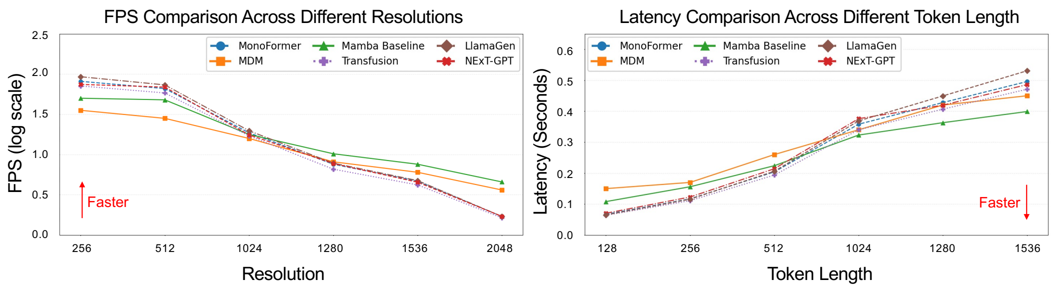


Figure 5. Comparison between Mamba Baseline, MonoFormer, and ours MDM on inference speed test. The left shows the inference speed of the model FPS at different resolutions. The right shows the inference speed of the model latency at different token lengths.

Model	Image/Text Scan Switch	FPS w log scale↑	FID↓
Model w Mamba	①②③④/①②	1.357	**2.49**
Model w Mamba	①②/①	1.405	3.96
Model w Transformer	-	**1.914**	6.72

Table 3. Ablation on ImageNet 256×256 image generation.

caption data. For the image generation results, we evaluated the FID and Gen Eval performance indicators of the model-generated images. MDM still achieved the top three performance levels and achieved SOTA on Gen Eval.

Text Generation. In the image-to-text task image captioning, according to the settings on image generation on the COCO dataset, we tested the caption data of the model based on the model outputting both text and image data using the CIDEr indicator. The results showed that MDM ranked second among all models, as shown in Tab. 2. While in task VQA, MDM achieves competitive performance, surpassing several traditional models including InstructBLIP, mPLUG-Owl, DREAMLLM, and Emu, although it still trails behind top-performing models in the field as shown in Tab. 2. In the text-to-text generation task, as shown in Tab. 2, MDM and the other end-to-end multi-modal models perform worse than well-known traditional models. This discrepancy may be attributed to the fact that these end-to-end models have some deviations in multimodal fusion and learning because they abandon multiple language encoders, visual encoders, and multimodal fusion encoders. However, when compared with the other two end-to-end models, MDM excels, outperforming MonoFormer and surpassing Chameleon on seven out of ten datasets.

5.3. Discussion

5.3.1. Performance Analysis

As demonstrated in Fig. 3, MDM shows the ability to generate image and text simultaneously in multiple rounds of dialogue and perform well in QA&VQA. Some results even exceed those of GPT-4V, particularly evident in the second and third rows of Fig. 4 which is a hybrid output process for the MDM model. Due to this, we set the model to generate corresponding images for the description text while simultaneously generating image captioning.

This enhanced performance stems from MDM's multi-step selection diffusion decoder, which leverages Mamba's integrated selection and denoising capabilities to maintain focused attention on both textual and visual details. Validating our complexity analysis in Appendix E, MDM demonstrates superior efficiency compared to end-to-end Transformer models when processing long sequences, as shown in Fig. 5, particularly outperforming other end-to-end multi-modal models for sequences exceeding 1280 tokens.

5.3.2. Ablations

Our ablation studies examine the impact of both the selection process and Mamba block components. Reducing the number of image/text scan switch sequences from 6 ('①②③④/①②') to 3 ('①②/①'), as shown in Tab. 3, improves inference speed but degrades image quality, as fewer scan switch sequences limit the model's ability to capture accurate information in complex sequences. Additionally, replacing the Mamba block with the Transformer further deteriorates output image quality, suggesting Mamba's temporal network architecture is better suited for representing diffusion relationships during the denoising process.

6. Conclusion

This paper introduces MDM (Multi-Modal Diffusion Mamba), a novel end-to-end architecture that significantly enhances multi-modal processing through two key innovations: a unified diffusion objective and an efficient selection mechanism leveraging Mamba's state-space structure. By integrating variational autoencoder with multi-step selection diffusion, MDM achieves SOTA overall performance in image generation and demonstrates remarkable versatility across various tasks, including image-to-text, text-to-text and text-image-to-text-image. Our comprehensive experiments illustrate that MDM consistently surpasses traditional end-to-end multi-modal models, particularly in processing high-resolution images and long-sequence text, while maintaining computational efficiency. The model's ability to unify different modalities under a single objective, coupled with its superior management of temporal relationships in the diffusion process, establishes a promising direction for future multi-modal architecture.

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3, 7

[3] Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient transformers. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 72–86. PMLR, 2023. 1

[4] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 1, 3

[5] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 1, 3

[6] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023. 7

[7] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7432–7439, 2020. 2, 6, 7

[8] GPTV System Card. Openai, 2023. 7

[9] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019. 2, 7

[10] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 2, 7

[11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 2, 7

[12] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6

[14] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6, 7

[16] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023. 6

[17] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 1, 3, 7

[18] Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Xuhui Liu, Juan Zhang, Guodong Guo, and Baochang Zhang. Fusion-mamba for cross-modality object detection. *arXiv preprint arXiv:2404.09146*, 2024. 1, 3

[19] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 4

[20] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 6, 7

[21] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 3, 6, 7

[22] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, Youqiang Zhang, and Junshi Huang. Dimba: Transformer-mamba diffusion models. *arXiv preprint arXiv:2406.01159*, 2024. 1

[23] 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. 1, 3

[24] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 6

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 6

[26] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 5

[27] Ahan Gupta, Yueming Yuan, Yanqi Zhou, and Charith Mendis. Flurka: Fast fused low-rank & kernel attention. *arXiv preprint arXiv:2306.15799*, 2023. 1

[28] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 2, 6

[29] Insu Han, Rajesh Jayaram, Amin Karbasi, Vahab Mirrokni, David P Woodruff, and Amir Zandieh. Hyperattention: Long-context attention in near-linear time. *arXiv preprint arXiv:2310.05869*, 2023. 1

[30] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024. 1

[31] Wanggui He, Siming Fu, Mushui Liu, Xierui Wang, Wenyi Xiao, Fangxun Shu, Yi Wang, Lei Zhang, Zhelun Yu, Haoyuan Li, et al. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. *arXiv preprint arXiv:2407.07614*, 2024. 1, 3

[32] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 2, 6, 7

[33] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 2, 7

[34] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 6

[36] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 6, 7

[37] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes S Fischer, and Björn Ommer. Zigma: A dit-style zigzag mamba diffusion model. *arXiv preprint arXiv:2403.13802*, 2024. 1

[38] 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. 6

[39] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 7

[40] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2, 6

[41] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 1, 3

[42] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 3, 4, 5, 6

[43] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[44] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[45] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. 4, 6

[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 3

[47] 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. 6, 7

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 6, 7

[49] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Lily Lee, Kaichen Zhou, Pengju An, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. *arXiv preprint arXiv:2406.04339*, 2024. 1, 3

[50] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 4

[51] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 5, 6

[52] 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. 6, 7

[53] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2, 6

[54] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new

dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018. 2, 6
[55] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 6
[56] Matteo Pagliardini, Daniele Paliotta, Martin Jaggi, and François Fleuret. Fast attention over long sequences with dynamic sparse flash attention. *Advances in Neural Information Processing Systems*, 36:59808–59831, 2023. 1
[57] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 6, 7
[58] 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. 1, 3, 6, 7
[59] Ramya Prabhu, Ajay Nayak, Jayashree Mohan, Ramachandran Ramjee, and Ashish Panwar. vattention: Dynamic memory management for serving llms without pagedattention. *arXiv preprint arXiv:2405.04437*, 2024. 1
[60] Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024. 1, 3
[61] Haohao Qu, Liangbo Ning, Rui An, Wenqi Fan, Tyler Derr, Xin Xu, and Qing Li. A survey of mamba. *arXiv preprint arXiv:2408.01129*, 2024. 1
[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3
[63] 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. 6, 7
[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 3, 6, 7
[65] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64 (9):99–106, 2021. 2, 6
[66] Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019. 1
[67] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6
[68] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 6
[69] 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. 1, 3, 6, 7
[70] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 7
[71] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1, 3, 6, 7
[72] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7
[73] Yao Teng, Yue Wu, Han Shi, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Dim: Diffusion mamba for efficient high-resolution image synthesis. *arXiv preprint arXiv:2405.14224*, 2024. 1, 2, 5, 6
[74] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 7
[75] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*, 2023. 1
[76] Zifu Wan, Pingping Zhang, Yuhao Wang, Silong Yong, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256*, 2024. 1, 3
[77] Xinghan Wang, Zixi Kang, and Yadong Mu. Text-controlled motion mamba: Text-instructed temporal grounding of human motion. *arXiv preprint arXiv:2404.11375*, 2024. 3
[78] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 1, 3, 6, 7
[79] 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. 6, 7
[80] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8239–8249, 2024. 1
[81] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9847–9857, 2021. 3

[82] Zhe Yang, Wenrui Li, and Guanghui Cheng. Shmamba: Structured hyperbolic state space model for audio-visual question answering. *arXiv preprint arXiv:2406.09833*, 2024. 1, 3

[83] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 7

[84] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2, 6

[85] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 6, 7

[86] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 3

[87] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019. 2, 6

[88] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024. 7

[89] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024. 1, 2, 3, 6, 7

[90] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 1, 3, 6, 7