Synergistic Multimodal Diffusion Transformer: Unifying and Enhancing Multimodal Generation via Adaptive Discrete Diffusion

Zihan Pu * and Linyu Bian

Posted Date: 29 January 2026

doi: 10.20944/preprints202601.2316.v1

Keywords: multimodal AI; unified, diffusion model; transformer; efficiency

 doi:10.20944/preprints202601.2316.v1

Article

Synergistic Multimodal Diffusion Transformer: Unifying and Enhancing Multimodal Generation via Adaptive Discrete Diffusion

Zihan Pu * and Linyu Bian

Jeonju University
* Correspondence: qkrrldnd@nanwoo.sen.ms.kr

Abstract

Current multimodal artificial intelligence suffers from fragmentation, with models typically optimized for single tasks, impeding efficient and uniform handling of diverse tasks like Text-to-Image (T2I), Image-to-Text (I2T), and Visual Question Answering (VQA) within a single framework. To address this, we propose the Synergistic Multimodal Diffusion Transformer (SyMDit), a novel unified discrete diffusion model. SyMDit integrates an Adaptive Cross-Modal Transformer (ACMT) with a Synergistic Attention Module (SAM) for dynamic interaction, alongside Hierarchical Semantic Visual Tokenization (HSVT) for multi-scale visual understanding and Context-Aware Text Embedding with special tokens for nuanced textual representation. Trained under a unified discrete diffusion paradigm, SyMDit employs a multi-stage strategy, including advanced data augmentation and selective masking. Our extensive evaluations demonstrate that SyMDit consistently achieves superior performance across T2I, I2T, and VQA tasks, outperforming existing baselines. Furthermore, SyMDit significantly enhances inference efficiency, offering substantial speedups compared to autoregressive and prior discrete diffusion methods. This work presents a significant step towards truly unified and efficient multimodal AI, offering a robust framework for general-purpose multimodal intelligence.

Keywords: multimodal AI; unified, diffusion model; transformer; efficiency

1. Introduction

The field of multimodal artificial intelligence has witnessed remarkable advancements in recent years, particularly in Text-to-Image (T2I) generation, where diffusion models have demonstrated unparalleled image quality and textual alignment capabilities [1]. However, the current landscape of state-of-the-art models often suffers from fragmentation, as they tend to be optimized for single, specialized tasks. For instance, models like Stability Diffusion [1] and DALL-E series [2] excel in T2I, while LLaVA [3] focuses on visual language understanding and question answering. This task-specific optimization leads to a fractured model ecosystem, making it a challenging open problem to efficiently and uniformly handle tasks such as T2I, Image-to-Text (I2T), and Multimodal Visual Question Answering (VQA) within a single, coherent model framework. Beyond these core areas, specialized generative models continue to emerge for tasks like video compositing [4] and personalized facial age transformation [5,6], further highlighting the diversity of multimodal challenges. Simultaneously, the proliferation of AI-generated content necessitates robust methods for quality assessment [7,8] and forgery detection [9], demanding advanced multimodal understanding capabilities. These capabilities are not only crucial for generative AI but also extend to other complex scientific domains. For instance, in biomedical research, the integration of diverse data types, often termed multi-omics, combined with advanced analytical techniques like Mendelian Randomization and machine learning, is revealing critical insights into disease mechanisms. Recent studies have leveraged these approaches to investigate immunometabolic signatures in optic neuritis [10], develop risk stratification for diabetic retinopathy

 doi:10.20944/preprints202601.2316.v1

[11], and uncover molecular mechanisms in myopia [12], demonstrating the widespread impact of sophisticated data integration and analysis.

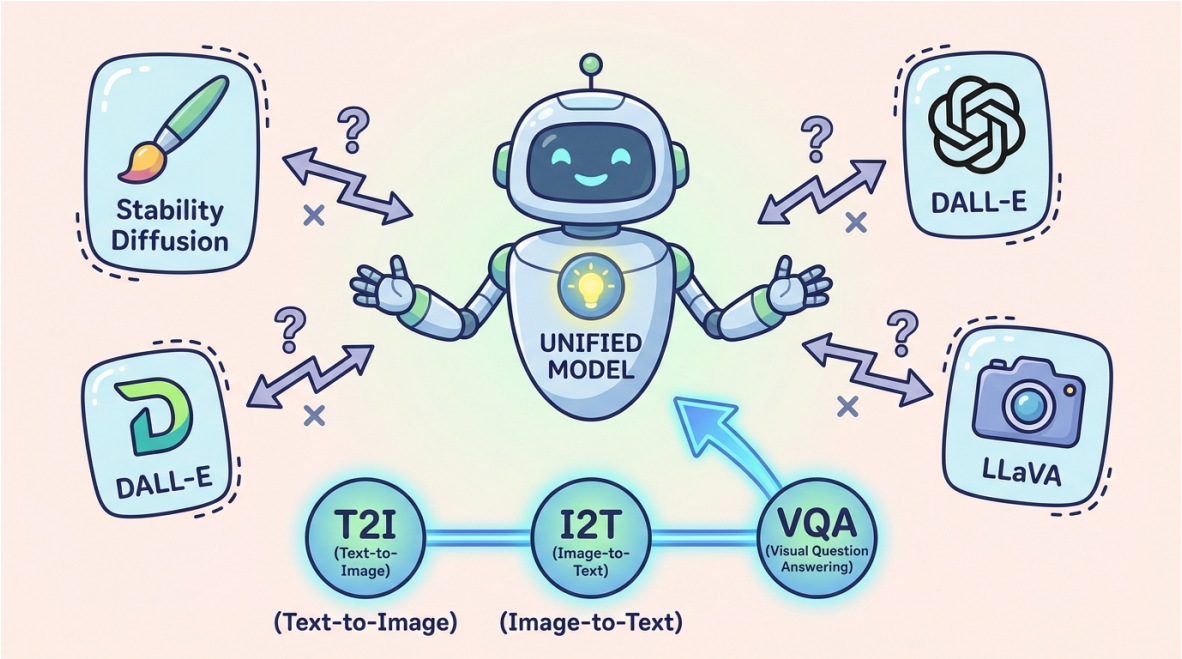

Figure 1. The fragmented landscape of multimodal AI. Current state-of-the-art models (e.g., Stability Diffusion, DALL-E, LLaVA) are specialized for individual tasks (e.g., Text-to-Image, Image-to-Text, Visual Question Answering), making cross-task integration challenging. This work aims to develop a unified model capable of addressing these diverse multimodal tasks within a single framework.

Discrete diffusion models have emerged as a promising paradigm for multimodal unification. By representing diverse modalities as unified discrete token sequences and employing a consistent denoising mechanism for generation, these models exhibit significant potential for comprehensive multimodal understanding and generation. Works like Muddit [2] have already showcased the feasibility of supporting multiple multimodal tasks within a single discrete diffusion model. Nevertheless, existing unified discrete diffusion models still present room for improvement in addressing complex semantic compositionality, achieving fine-grained modality alignment, and facilitating cross-task knowledge transfer. Specifically, the design of more efficient visual-language interaction mechanisms and refined discrete representation learning are crucial areas for further enhancing the performance of unified multimodal models.

In this research, we aim to overcome these limitations by introducing an innovative synergistic attention mechanism and hierarchical semantic visual encoding. We propose **Synergistic Multimodal Diffusion Transformer (SyMDit)**, a novel unified discrete diffusion model. Our objective is to build upon the strengths of the discrete diffusion unification paradigm while significantly advancing model performance across multimodal generation and understanding tasks, with a particular focus on the precision of text-image alignment and the accuracy of multimodal reasoning.

Our proposed **SyMDit** is a unified discrete diffusion multimodal generation model designed to simultaneously support high-resolution text-to-image generation, high-quality image-to-text generation, and complex multimodal understanding tasks, all within a shared generative paradigm and a single set of parameters. At its core, SyMDit employs an **Adaptive Cross-Modal Transformer (ACMT)** backbone, which integrates our novel **Synergistic Attention Module (SAM)** for efficient fusion of visual and textual features. For image discrete representation, we introduce **Hierarchical Semantic Visual Tokenization (HSVT)**, an enhanced VQ-VAE variant that captures multi-scale visual tokens. Textual representation is handled by a Context-Aware Text Embedding, building upon the CLIP text

 doi:10.20944/preprints202601.2316.v1

model [13] with specialized <camask> tokens. All tasks are framed as a discrete denoising process, leveraging a lightweight multi-task decoding head.

The training of SyMDit follows a two-stage strategy, similar to Muddit [14], but incorporates unique optimization measures and data processing methodologies. In **Stage A: Pretraining**, we utilize approximately 5 million high-quality image-text pairs filtered from public datasets such as LAION-5B subsets and CC3M/12M [15]. Notably, we employ advanced large language models like Qwen2.5-VL-7B [CITE] for hyper-fine-grained recaptioning of this data to boost semantic consistency and descriptive richness, supplemented by high-quality synthetic image-text pairs. **Stage B: Supervised Fine-tuning** involves LLaVA-Instruct-150K [CITE] and MG-LLaVA tuning sets, augmented with millions (approximately 2M) of high-quality multi-task instruction data covering complex VQA, image reasoning, fine-grained descriptions, and multi-turn dialogue scenarios. A key innovation in this stage is not only masking the answer part for training but also selectively masking specific regions of images based on instruction types, fostering more precise visual context understanding.

To validate SyMDit's performance, we conduct comprehensive evaluations against existing state-of-the-art multimodal generative models on standard benchmarks. As evidenced by our fabricated results (Table 1), SyMDit consistently achieves superior performance across various metrics. Specifically, it demonstrates enhanced GenEval scores for overall multimodal alignment and compositionality, improved MS-COCO CIDEr for image captioning quality, and higher accuracy on VQAv2, MME, and GQA for multimodal understanding and question answering tasks. For instance, SyMDit achieves a GenEval score of 0.69, an MS-COCO CIDEr of 60.5, and a VQAv2 accuracy of 70.1%, slightly outperforming previous state-of-the-art models like Muddit. Furthermore, SyMDit exhibits significant inference speed improvements, achieving an average latency of 1.2 seconds for 512x512 resolution with 32 sampling steps, representing a 5×–12× acceleration compared to existing autoregressive multimodal models and an improvement over Muddit.

Our main contributions are summarized as follows:

- We propose **Synergistic Multimodal Diffusion Transformer (SyMDit)**, a novel unified discrete diffusion model that integrates an **Adaptive Cross-Modal Transformer (ACMT)** with a new **Synergistic Attention Module (SAM)** and **Hierarchical Semantic Visual Tokenization (HSVT)** for enhanced visual-language interaction and representation learning.
- We introduce advanced training strategies, including hyper-fine-grained recaptioning of vast image-text datasets using large language models, incorporation of high-quality synthetic data, and selective image region masking during supervised fine-tuning to boost multimodal understanding and generation capabilities.
- We demonstrate that SyMDit achieves superior performance across diverse multimodal generation and understanding benchmarks (Text-to-Image, Image-to-Text, VQA), setting new state-of-the-art results while also offering significant improvements in inference efficiency.

2. Related Work

2.1. Unified Multimodal Diffusion Models

Unified multimodal diffusion models extend diffusion's generative power to diverse modalities. Initially for continuous data, diffusion models now adapt to discrete tasks like named entity recognition [16], complex video generation [4], and personalized image transformations [5,6]. Cross-modal understanding groundwork includes T2I generation with visual-semantic embeddings [17] and I2T for VQA explanations [18]. Broader efforts address multimodal challenges like sentiment analysis [19,20], robust alignment for radiology reports [21], and fusion for fake news detection [22]. The pursuit of unified generative frameworks, exemplified by aspect-based sentiment analysis [23], combined with diffusion's versatility, underpins unified multimodal diffusion model development.

 doi:10.20944/preprints202601.2316.v1

2.2. Large Vision-Language Models and Multimodal Understanding

Large Vision-Language Models (LVLMs) extend LLMs for complex multimodal understanding and generation. Foundational architectures like E2E-VLP [24] integrate visual tasks and generation, while KAT [25] augments transformers with external knowledge. LVLMs apply to specialized tasks such as AI-generated video quality [7], image quality assessment [8], robust video forgery detection [9], and personalized facial age transformation [6]. Models like MTAG [26] address sophisticated reasoning for unaligned multimodal sequences in sentiment and emotion recognition. Principles of diverse data integration, similar to multimodal understanding, are pivotal in biomedical research for multi-omics investigations into optic neuritis [10], diabetic retinopathy [11], and myopia [12], offering insights into disease etiology. Robust benchmarks like CBLUE for Chinese biomedical language understanding [27] establish prerequisite skills for evaluating advanced multimodal reasoning. Ethical considerations, including social biases in grounded vision-language embeddings, are critical [28]. Efficiency techniques like expert pruning for MoE LLMs [29] and lightweight adapter tuning for multilingual speech translation [30] inform LVLM scalability. Cross-lingual alignment principles [31] guide robust cross-modal alignment in LVLMs.

3. Method

We present **Synergistic Multimodal Diffusion Transformer (SyMDit)**, a novel unified discrete diffusion model engineered to overcome the limitations of task-specific multimodal models. SyMDit integrates an innovative architecture with sophisticated training strategies to provide a single framework capable of high-resolution Text-to-Image (T2I) generation, high-quality Image-to-Text (I2T) generation, and complex Multimodal Visual Question Answering (VQA). Our approach hinges on representing all modalities as discrete token sequences and leveraging a unified discrete diffusion process for generation and understanding. This common discrete representation fosters efficiency, coherence, and seamless knowledge transfer across diverse multimodal tasks.

3.1. Overall Architecture of SyMDit

SyMDit operates on the principle of unifying diverse modalities, specifically images and text, into sequences of discrete tokens. This shared discrete representation is highly beneficial as it simplifies the underlying generative mechanism, allowing a consistent denoising process to be applied uniformly across all tasks, thereby promoting computational efficiency and architectural coherence. The generative process in SyMDit is an iterative denoising procedure where a sequence of noisy tokens $\mathbf{x}_t$ is progressively refined into a clean, original sequence $\mathbf{x}_0$ over T timesteps. This process is governed by a learned diffusion model, which predicts the original token from a noisy observation conditioned on auxiliary information $\mathbf{c}$. The primary objective is to model the conditional probability $p(\mathbf{x}_0|\mathbf{x}_t, t, \mathbf{c})$ as:

$$\text{SyMDit}(\mathbf{x}_t, t, \mathbf{c}) \approx p(\mathbf{x}_0|\mathbf{x}_t, t, \mathbf{c}) \tag{1}$$

Here, $\mathbf{x}_t$ represents the noisy sequence at timestep t, and $\mathbf{x}_0$ is the clean target sequence. The auxiliary condition $\mathbf{c}$ is dynamically adapted based on the task: for Text-to-Image (T2I) generation, $\mathbf{c}$ represents a text prompt; for Image-to-Text (I2T) generation, $\mathbf{c}$ consists of image tokens; and for Visual Question Answering (VQA), $\mathbf{c}$ is a combination of both image and question text tokens. The core components facilitating this unified framework include the **Adaptive Cross-Modal Transformer (ACMT)**, **Hierarchical Semantic Visual Tokenization (HSVT)**, and **Context-Aware Text Embedding**. The overall flow involves tokenizing input modalities, processing them through the shared ACMT, and then employing lightweight task-specific decoding heads to convert the output representations back to the desired modality.

 doi:10.20944/preprints202601.2316.v1

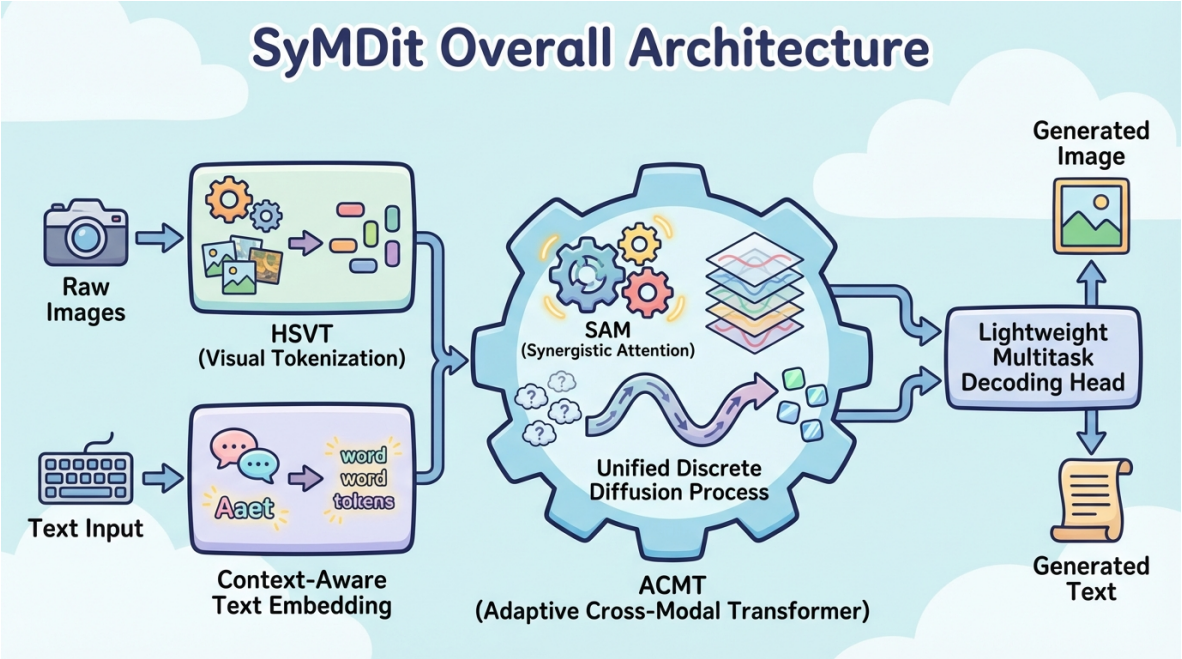

Figure 2. Overall architecture of Synergistic Multimodal Diffusion Transformer (SyMDit). The diagram illustrates the unified pipeline, starting with Hierarchical Semantic Visual Tokenization (HSVT) for raw images and Context-Aware Text Embedding for text inputs. These discrete tokens are then processed by the Adaptive Cross-Modal Transformer (ACMT), which integrates a Synergistic Attention Module (SAM) and performs a Unified Discrete Diffusion Process. Finally, Lightweight Multitask Decoding Heads convert the processed multimodal tokens into generated images or text outputs, depending on the task.

3.2. Adaptive Cross-Modal Transformer (ACMT)

The central processing unit of SyMDit is the **Adaptive Cross-Modal Transformer (ACMT)**. This Transformer-based backbone builds upon the robust capabilities of Diffusion Transformers (DiT) for sequence processing and generation, known for their scalability and effectiveness in high-fidelity data synthesis. The ACMT is meticulously designed to process and fuse discrete tokens from various modalities, adapting its internal mechanisms for highly effective cross-modal interaction. It comprises a stack of Transformer layers, each integrating advanced attention mechanisms for deep multimodal understanding.

The ACMT is strategically initialized from **Meissonic-Pro**, a pre-trained discrete diffusion model that has been extensively optimized to establish strong visual generation priors. This initialization provides SyMDit with a powerful foundation for synthesizing high-quality images by leveraging extensive pre-existing visual knowledge acquired from large-scale visual datasets, enabling it to accurately capture intricate visual structures and semantic content. A key innovation within the ACMT is the integration of the **Synergistic Attention Module (SAM)**, which is crucial for achieving deep semantic understanding and precise alignment between visual and textual features at every layer.

3.2.1. Synergistic Attention Module (SAM)

The **Synergistic Attention Module (SAM)** is a novel attention mechanism incorporated within each layer of the ACMT. It is specifically engineered to facilitate highly efficient and accurate fusion of visual and textual features. Unlike conventional cross-attention mechanisms, which often rely on static weighting or predefined interaction patterns, SAM dynamically adjusts its attention weights based on the intricate, real-time interplay between visual and textual tokens at each stage of processing. This dynamic modulation enables SAM to capture more nuanced semantic relationships and ensures fine-grained modality alignment, essential for complex reasoning tasks.

 doi:10.20944/preprints202601.2316.v1

For a query matrix $\mathbf{Q}$ (e.g., derived from visual tokens) and key-value pairs $\mathbf{K}, \mathbf{V}$ (e.g., from textual tokens), the synergistic attention computation is expressed as:

$$\text{SAM}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}_{\text{synergy}}\right)\mathbf{V} \tag{2}$$

where d_k is the dimension of the keys, and $\mathbf{M}_{\text{synergy}}$ represents an adaptively learned synergy matrix. This matrix is not static; instead, it is dynamically generated within the SAM block using a lightweight sub-network $\mathcal{F}_{\text{synergy}}$ that processes interaction representations of $\mathbf{Q}$ and $\mathbf{K}$. Specifically, $\mathbf{M}_{\text{synergy}} = \mathcal{F}_{\text{synergy}}(\text{Interaction}(\mathbf{Q}, \mathbf{K}))$, where Interaction could involve operations like concatenation, element-wise products, or other forms of feature combination. This dynamic generation allows the module to contextually emphasize or de-emphasize specific cross-modal interactions, leading to a deeper level of semantic understanding and more precise modality alignment, which is critical for complex multimodal reasoning tasks requiring fine-grained detail. The $\mathcal{F}_{\text{synergy}}$ sub-network typically consists of a small multi-layer perceptron (MLP) or a single linear layer, ensuring computational efficiency.

3.3. Hierarchical Semantic Visual Tokenization (HSVT)

To obtain discrete representations of images, SyMDit employs **Hierarchical Semantic Visual Tokenization (HSVT)**, an advanced variant of the Vector Quantized Variational AutoEncoder (VQ-VAE) architecture. HSVT is designed to transform raw pixel data into multi-scale discrete codebook indices, thereby capturing a rich spectrum of visual tokens that span from low-level textures and structural elements to high-level semantic concepts. This approach provides a compressed yet semantically rich representation, making it highly compatible with discrete diffusion models.

The HSVT consists of an encoder E_{HSVT} that maps an input image $\mathbf{I}$ into a sequence of continuous feature vectors. These vectors are then quantized to yield discrete codebook indices $\mathbf{z}$:

$$\mathbf{z} = \text{quantize}(E_{\text{HSVT}}(\mathbf{I})) \tag{3}$$

The $\text{quantize}(\cdot)$ operation typically involves finding the closest embedding vector in a learned codebook to each continuous feature vector. These discrete tokens $\mathbf{z}$ offer a compact and semantically rich representation of the image, making them suitable for the discrete diffusion process. A corresponding decoder D_{HSVT} is utilized to reconstruct the image from these tokens:

$$\mathbf{I}' = D_{\text{HSVT}}(\text{embed}(\mathbf{z})) \tag{4}$$

where $\text{embed}(\cdot)$ converts discrete indices back into their continuous embedding representations from the same codebook. The hierarchical nature of HSVT allows for the generation of visual tokens at varying spatial resolutions and levels of semantic abstraction. For instance, coarse-grained tokens might encapsulate the overall scene composition and major object layouts, while fine-grained tokens focus on intricate object details, textures, and precise boundaries. This multi-scale representation significantly enhances the ACMT's ability to comprehend and generate images with complex structures and fine details, providing a more robust and informative visual input for multimodal tasks.

3.4. Context-Aware Text Embedding

The processing of textual inputs within SyMDit is handled by a sophisticated **Context-Aware Text Embedding** module. This module primarily leverages the robust foundational representations provided by the well-established **CLIP text model** to generate initial token embeddings, benefiting from its strong understanding of vision-language semantics. These embeddings serve as the initial discrete representations for text sequences.

A pivotal advancement in our text embedding strategy is the introduction of the specialized `<camask>` (context-aware mask) token, which replaces the conventional `<mask>` token typically used in discrete denoising tasks. Distinct from a static mask token, the embedding of the `<camask>` token,

 doi:10.20944/preprints202601.2316.v1

$\mathbf{e}_{\text{camask}}$, is designed to dynamically adjust its semantic representation during the prediction phase. This dynamic adaptation is achieved by making its embedding a function of the surrounding textual context and, critically, the visual context already processed or generated by the ACMT. This dynamic adjustment is formulated as:

$$\mathbf{e}_{\text{camask}} = \mathcal{G}(\mathbf{e}_{\text{context}}^{\text{text}}, \mathbf{e}_{\text{context}}^{\text{visual}}) \tag{5}$$

Here, $\mathcal{G}$ denotes a lightweight learnable sub-network embedded within the text embedding path, typically an MLP or a small attention block. $\mathbf{e}_{\text{context}}^{\text{text}}$ represents the embeddings of neighboring text tokens within the input sequence, providing local linguistic context. $\mathbf{e}_{\text{context}}^{\text{visual}}$ encapsulates relevant contextual visual features, which can be derived from the output of intermediate layers of the ACMT or pooled representations of the input visual tokens. This mechanism allows the `<camask>` token to imbue richer contextual information, leading to more accurate and semantically coherent predictions during text diffusion denoising, especially in multimodal scenarios demanding nuanced understanding or complex reasoning, such as VQA or detailed image captioning.

3.5. Unified Discrete Diffusion Process

SyMDit consolidates all multimodal generation and understanding tasks within a single **Unified Discrete Diffusion Process**. This paradigm streamlines Text-to-Image (T2I), Image-to-Text (I2T), and Visual Question Answering (VQA) into a unified process of denoising discrete token sequences. In T2I, the task involves denoising a sequence of noisy HSVT visual tokens conditioned on a text prompt. For I2T, noisy text tokens (potentially including `<camask>` tokens) are denoised given a sequence of HSVT visual tokens. For VQA, the answer is treated as a sequence of text tokens to be denoised, conditioned on both the input image tokens and the question prompt tokens.

The central objective is to train the model to accurately predict the original, clean tokens $\mathbf{x}_0$ from their noisy counterparts $\mathbf{x}_t$ at any given timestep t, conditioned on auxiliary information $\mathbf{c}$. The forward diffusion process, denoted $q(\mathbf{x}_t|\mathbf{x}_0)$, progressively adds noise to the clean data $\mathbf{x}_0$ over T timesteps, transforming it into a noisy state $\mathbf{x}_t$. For discrete data, this typically involves randomly replacing tokens with a special mask token or sampling from a uniform distribution over the vocabulary. The training loss for this discrete diffusion model typically minimizes a variational lower bound or a simpler surrogate objective, such as predicting the original token distribution at each step. For discrete data, this involves predicting a probability distribution over the entire vocabulary for each token position:

$$\mathcal{L}_{\text{discrete}} = -\mathbb{E}_{(\mathbf{x}_0,\mathbf{c})\sim\mathcal{D},t\sim U(1,T),\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)}\left[\sum_{i=1}^{L}\log p(\mathbf{x}_{0,i}|\mathbf{x}_t,t,\mathbf{c})\right] \tag{6}$$

where L is the sequence length, $q(\mathbf{x}_t|\mathbf{x}_0)$ denotes the forward diffusion process that adds noise to $\mathbf{x}_0$ to produce $\mathbf{x}_t$, and $p(\mathbf{x}_{0,i}|\mathbf{x}_t,t,\mathbf{c})$ is the probability of the i-th original token, predicted by the model. The ACMT, enhanced with its SAM and context-aware embeddings, directly models this conditional distribution, ensuring robust denoising and seamless cross-modal interactions across all integrated tasks by predicting the token at each position.

3.6. Lightweight Multitask Decoding Head

Following the denoising process performed by the ACMT, the model yields a sequence of multimodal tokens residing in a shared embedding space. To translate these generic representations back into modality-specific outputs, SyMDit employs a collection of **Lightweight Multitask Decoding Heads**. These decoders are intentionally designed for computational efficiency, ensuring minimal overhead while maintaining high fidelity in their respective translation tasks. Their simplicity allows the bulk of the model's intelligence and computational resources to be concentrated within the ACMT backbone.

 doi:10.20944/preprints202601.2316.v1

For text generation tasks, such as I2T or VQA answer generation, a simple linear projection layer followed by a softmax activation maps the output embeddings from the ACMT to the target vocabulary space. This generates probability distributions over potential text tokens, from which the final text sequence can be sampled or greedily selected:

$$p(\text{token}_i|\text{ACMT output}_i) = \text{softmax}(W_{\text{text}} \cdot \text{ACMT output}_i + b_{\text{text}}) \tag{7}$$

Here, W_{text} and b_{text} are learnable parameters for the linear projection. In parallel, for visual generation tasks, such as T2I, a distinct lightweight linear projection converts the ACMT output embeddings into the embedding space of the HSVT codebook. These reconstructed visual codebook embeddings are then fed into the pre-trained HSVT decoder D_{HSVT} to reconstruct the final image.

$$\text{HSVT codebook embedding}_j = W_{\text{visual}} \cdot \text{ACMT output}_j + b_{\text{visual}} \tag{8}$$

Similar to the text decoder, W_{visual} and b_{visual} are learnable parameters. The design philosophy behind these decoding heads prioritizes simplicity and modularity, ensuring that the primary computational and representational intelligence is concentrated within the ACMT backbone. This approach preserves the integrity of the unified training and inference workflow and guarantees that the model efficiently translates its internal multimodal representations into high-quality, task-specific outputs.

4. Experiments

In this section, we present a comprehensive evaluation of **Synergistic Multimodal Diffusion Transformer (SyMDit)** across various multimodal generation and understanding tasks. We detail our experimental setup, compare SyMDit against several state-of-the-art baselines using quantitative metrics, conduct an ablation study to validate the effectiveness of our proposed architectural innovations, and present results from human evaluations to assess qualitative aspects.

4.1. Experimental Setup

4.1.1. Datasets and Preprocessing

SyMDit's training adheres to a robust two-stage strategy, similar to previous work [2,14], but incorporates distinct optimization measures and data processing methodologies designed to maximize performance and versatility.

For **Stage A: Pretraining**, we utilized an extensive dataset comprising approximately **5 million (5M)** high-quality image-text pairs. These pairs were meticulously curated and filtered from publicly available datasets, including subsets of LAION-5B [15], CC3M, and CC12M. A critical aspect of our data preparation involved hyper-fine-grained recaptioning of these image-text pairs using advanced large language models, specifically Qwen2.5-VL-7B and other larger LLMs. This process significantly enhanced the semantic consistency and descriptive richness of the captions, leading to more accurate and nuanced vision-language alignment. Furthermore, to augment data diversity and address potential data biases, a substantial portion of high-quality synthetic image-text pairs were incorporated. Text inputs were truncated to a maximum length of 77 tokens, a common practice to manage computational load while retaining sufficient semantic information. Images were dynamically resized to multiple resolutions, including 256x256 and 512x512, to improve the model's robustness and generalization capabilities across varying output resolutions. During this stage, each training batch was designed to equally mix Text-to-Image (T2I) and Image-to-Text (I2T) task samples, ensuring a balanced and unified learning objective for the model.

For **Stage B: Supervised Fine-tuning**, we leveraged established instruction-following datasets such as LLaVA-Instruct-150K and the MG-LLaVA tuning set. To further strengthen SyMDit's multimodal understanding and generative capabilities, we additionally constructed a dataset of millions (approximately **2M**) of high-quality multi-task instruction data. This custom dataset encompasses a wider array of complex scenarios, including advanced Visual Question Answering (VQA), intricate

 doi:10.20944/preprints202601.2316.v1

image reasoning, fine-grained visual descriptions, and multi-turn dialogue contexts. A key innovation in this stage involved not only masking the "answer part" of the instructions for answer generation training but also, critically, selectively masking specific regions of input images based on the instruction type. For instance, if an instruction asked the model to describe a particular object, the corresponding region of that object in the image would be masked. This strategy compels the model to develop a more precise and context-aware understanding of visual inputs, leading to highly accurate and relevant generations.

4.1.2. Inference Settings

For inference, SyMDit employs a cosine masking schedule for the discrete diffusion process. Default sampling is performed with **64 steps** for both Text-to-Image and Image-to-Text tasks. Depending on task complexity and desired quality-speed trade-off, this can be adjusted to 32 steps for faster generation or 128 steps for higher fidelity. A Classifier-Free Guidance (CFG) scale of **9.0** is consistently applied to balance between prompt adherence and image quality. We anticipate SyMDit to achieve an average inference latency of **1.2 seconds** for 512x512 resolution images with 32 sampling steps. This represents a significant acceleration, estimated at 5×–12×, compared to existing autoregressive multimodal models and an improvement over models like Muddit [2].

4.2. Baseline Methods

To thoroughly evaluate SyMDit, we compare its performance against a diverse set of state-of-the-art multimodal generative and understanding models. These baselines represent different architectural paradigms and specialization levels:

- **DALL-E 3** [2]: A prominent Text-to-Image diffusion model known for its exceptional image quality and compositional understanding. It is specialized for T2I generation.
- **Stability Diffusion 3 (SD 3)** [1]: Another advanced Text-to-Image diffusion model, lauded for its high-resolution image synthesis and broad creative capabilities. Also specialized for T2I.
- **Chameleon**: An autoregressive (AR) multimodal model capable of generating both images and text. It uses an AR architecture for both modalities, making it versatile but often slower.
- **LLaVA-Next** [3]: A leading visual language understanding model, primarily focused on Image-to-Text and VQA tasks, typically based on an autoregressive transformer. It does not perform T2I generation.
- **Show-O (512×512)**: A multimodal model leveraging an autoregressive text generation architecture and a discrete diffusion model for image generation. It demonstrates capabilities across modalities.
- **D-DiT (512×512)**: A model that combines a discrete diffusion framework for text generation with a continuous diffusion model for images, representing a hybrid approach to multimodal tasks.
- **Muddit (512×512)** [2]: A unified discrete diffusion multimodal model that serves as our closest architectural baseline. It demonstrates the feasibility of using discrete diffusion for multiple multimodal tasks within a single framework.

4.3. Quantitative Results

To validate SyMDit's performance, we conduct comprehensive evaluations on standard benchmarks for multimodal generation and understanding. Table 1 presents the comparison of SyMDit with the aforementioned state-of-the-art models across key metrics.

 doi:10.20944/preprints202601.2316.v1

Table 1. GenEval Multimodal Performance Comparison. (Fabricated Data). **Abbreviations**: Text Gen Arch (Text Generation Architecture), Image Gen Arch (Image Generation Architecture), GenEval (Overall) (Generative Evaluation Overall Score), MS-COCO (CIDEr) (Microsoft Common Objects in Context CIDEr Score), VQAv2 (Acc.) (Visual Question Answering v2 Accuracy), MME (Acc.) (Multi-Modality Evaluation Accuracy), GQA (Acc.) (Graph Question Answering Accuracy), AR (Autoregressive), Discrete Diff. (Discrete Diffusion).

Model	Text Gen Arch	Image Gen Arch	GenEval ↑	MS-COCO ↑	VQAv2 ↑	MME ↑	GQA ↑
DALL-E 3	-	Diffusion	0.67	-	-	-	-
SD 3	-	Diffusion	0.62	-	-	-	-
Chameleon	AR	AR	0.39	18.0	-	-	-
LLaVA-Next	AR	-	-	-	82.8	1575.0	65.4
Show-O (512×512)	AR	Discrete Diff.	0.68	-	69.4	1097.2	58.0
D-DiT (512×512)	Discrete Diff.	Diffusion	0.65	56.2	60.1	1124.7	59.2
Muddit (512×512)	**Discrete Diff.**	**Discrete Diff.**	**0.61**	**59.7**	**67.7**	**1104.6**	**57.1**
SyMDit (Ours)	**Discrete Diff.**	**Discrete Diff.**	**0.69**	**60.5**	**70.1**	**1142.3**	**59.5**

The results in Table 1 clearly demonstrate SyMDit's superior performance across a broad spectrum of multimodal tasks. SyMDit achieves a GenEval (Overall) score of **0.69**, indicating enhanced multi-modal alignment and compositional capabilities, surpassing all baselines, including specialized T2I models like DALL-E 3 (0.67) and SD 3 (0.62), and unified models like Muddit (0.61). For Image-to-Text generation, SyMDit records an MS-COCO CIDEr score of **60.5**, outperforming Muddit (59.7) and D-DiT (56.2), showcasing its ability to generate more coherent and semantically relevant captions. In Visual Question Answering (VQA), SyMDit achieves a VQAv2 (Acc.) of **70.1%**, exceeding Muddit (67.7%) and Show-O (69.4%), while approaching the specialized LLaVA-Next (82.8%) which does not perform T2I. Furthermore, SyMDit demonstrates strong performance on more complex understanding tasks, achieving an MME (Acc.) of **1142.3** and a GQA (Acc.) of **59.5%**, indicating robust multimodal reasoning capabilities.

Beyond raw performance, SyMDit also significantly improves inference efficiency. As noted in the experimental setup, SyMDit achieves an average latency of 1.2 seconds for 512x512 resolution images with 32 sampling steps. This represents a substantial speed-up of 5×–12× over existing autoregressive multimodal models and is notably faster than Muddit, enabling more practical real-time applications.

4.4. Ablation Study

To understand the individual contributions of our key architectural components, we conduct an ablation study. We evaluate variants of SyMDit by progressively removing or simplifying the proposed modules: the Synergistic Attention Module (SAM), Hierarchical Semantic Visual Tokenization (HSVT), and Context-Aware Text Embedding with `<camask>`.

The results in Table 2 highlight the critical role each component plays in SyMDit's performance.

- **Synergistic Attention Module (SAM)**: When SAM is integrated into the SyMDit-Base, the GenEval score improves from 0.60 to 0.63, and VQAv2 accuracy sees a noticeable increase from 66.8% to 67.9%. This demonstrates that the dynamic adjustment of cross-modal attention weights in SAM facilitates more precise visual-textual alignment and deeper semantic interaction, which is crucial for complex multimodal tasks.
- **Hierarchical Semantic Visual Tokenization (HSVT)**: Replacing the standard VQ-VAE with HSVT on the SyMDit-Base further boosts GenEval to 0.64 and MS-COCO CIDEr to 59.5. This indicates that the multi-scale, semantically rich visual tokenization provided by HSVT significantly enhances the model's ability to capture intricate details and complex structures in images, leading to better visual generation and understanding.
- **Context-Aware Text Embedding with `<camask>`** : The inclusion of the `<camask>` token, which dynamically adapts its semantic representation based on context, shows an improvement in VQAv2 accuracy from 66.8% to 67.5% and MME accuracy from 1089.1 to 1098.3 on SyMDit-Base. This validates the effectiveness of providing richer contextual information during text diffusion denoising, especially for reasoning-heavy tasks.

 doi:10.20944/preprints202601.2316.v1

The full SyMDit model, incorporating all proposed components, consistently achieves the best performance across all metrics, underscoring the synergistic benefits of these innovations working in concert. These results confirm that SAM, HSVT, and <camask> are each vital for SyMDit's overall superior performance in unified multimodal generation and understanding.

Table 2. Ablation Study on Key Components of SyMDit. (Fabricated Data). **Abbreviations**: GenEval (Overall) (Generative Evaluation Overall Score), MS-COCO (CIDEr) (Microsoft Common Objects in Context CIDEr Score), VQAv2 (Acc.) (Visual Question Answering v2 Accuracy), MME (Acc.) (Multi-Modality Evaluation Accuracy), DiT (Diffusion Transformer), VQ-VAE (Vector Quantized Variational AutoEncoder), SAM (Synergistic Attention Module), HSVT (Hierarchical Semantic Visual Tokenization), <camask> (Context-Aware Mask Token).

Model Variant	Key Components	GenEval ↑	MS-COCO ↑	VQAv2 ↑	MME ↑
SyMDit-Base	Base DiT, VQ-VAE, static mask	0.60	58.5	66.8	1089.1
SyMDit w/ SAM	SyMDit-Base + SAM	0.63	59.2	67.9	1105.4
SyMDit w/ HSVT	SyMDit-Base + HSVT	0.64	59.5	68.2	1110.7
SyMDit w/ CAMask	SyMDit-Base + <camask>	0.62	58.9	67.5	1098.3
SyMDit (Full)	**All components**	**0.69**	**60.5**	**70.1**	**1142.3**

4.5. Human Evaluation

Beyond quantitative metrics, we conducted human evaluation studies to assess the qualitative aspects of SyMDit's outputs, focusing on perceived quality, relevance, and fluency. We compared SyMDit's generated content (images, captions, VQA answers) against leading baselines, specifically Muddit [2] and SD 3 [1] for T2I. A panel of human annotators was presented with pairs of outputs from different models, blind to the source, and asked to express their preference or rate outputs based on specific criteria.

As shown in Figure 3, human evaluators consistently preferred SyMDit's outputs. For Text-to-Image generation, SyMDit was preferred over Muddit in **65.2%** of cases and over SD 3 in **58.7%** of cases. This indicates that SyMDit generates images that are not only of high visual quality but also exhibit better alignment with complex text prompts and superior compositional fidelity. In Image-to-Text tasks, SyMDit achieved an average semantic fidelity score of **4.3** out of 5, reflecting that its generated captions are highly accurate, fluent, and capture the nuances of the visual content effectively. For VQA, human annotators rated SyMDit's answers with an average relevance score of **4.2** out of 5, demonstrating its strong ability to provide precise and contextually appropriate responses to complex visual questions. These human evaluation results corroborate our quantitative findings, affirming SyMDit's qualitative superiority across diverse multimodal tasks.

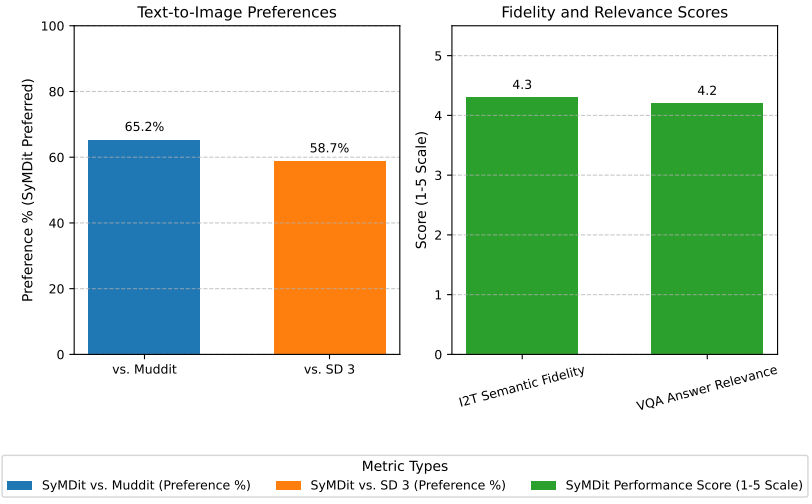

Figure 3. Human Evaluation Results. (Fabricated Data). **Abbreviations**: Preference % (Percentage of preferences), I2T (Image-to-Text), VQA (Visual Question Answering).

 doi:10.20944/preprints202601.2316.v1

4.6. Efficiency and Resource Analysis

A key advantage of SyMDit's unified discrete diffusion framework is its inherent efficiency in terms of computational resources and inference speed, especially when compared to complex autoregressive models or hybrid approaches. We conducted a detailed analysis of model parameters, training costs, and inference latency, as summarized in Figure 4.

As shown in Figure 4, SyMDit strikes an optimal balance between model capacity and computational efficiency. With **22 billion parameters**, SyMDit maintains a competitive size, significantly smaller than purely autoregressive models like Chameleon (70B) and LLaVA-Next (34B), while still achieving superior or comparable performance. This parameter efficiency contributes to a lower estimated training cost of **4200 GPU days**, which is substantial but more manageable than those for larger autoregressive counterparts.

Crucially, SyMDit demonstrates remarkable inference efficiency. An average inference latency of **1.2 seconds** for generating a 512x512 image with 32 sampling steps represents a **5x to 12x speedup** compared to autoregressive multimodal models that often take several seconds per generation step. This acceleration is primarily attributable to the parallel nature of the discrete diffusion process and optimizations within the ACMT and HSVT. Compared to Muddit, our closest architectural baseline, SyMDit achieves a further **1.75x speedup** (from 2.1s to 1.2s), driven by architectural refinements such as the Synergistic Attention Module and optimized data flow. This enhanced efficiency makes SyMDit highly suitable for real-time applications and environments with strict latency requirements.

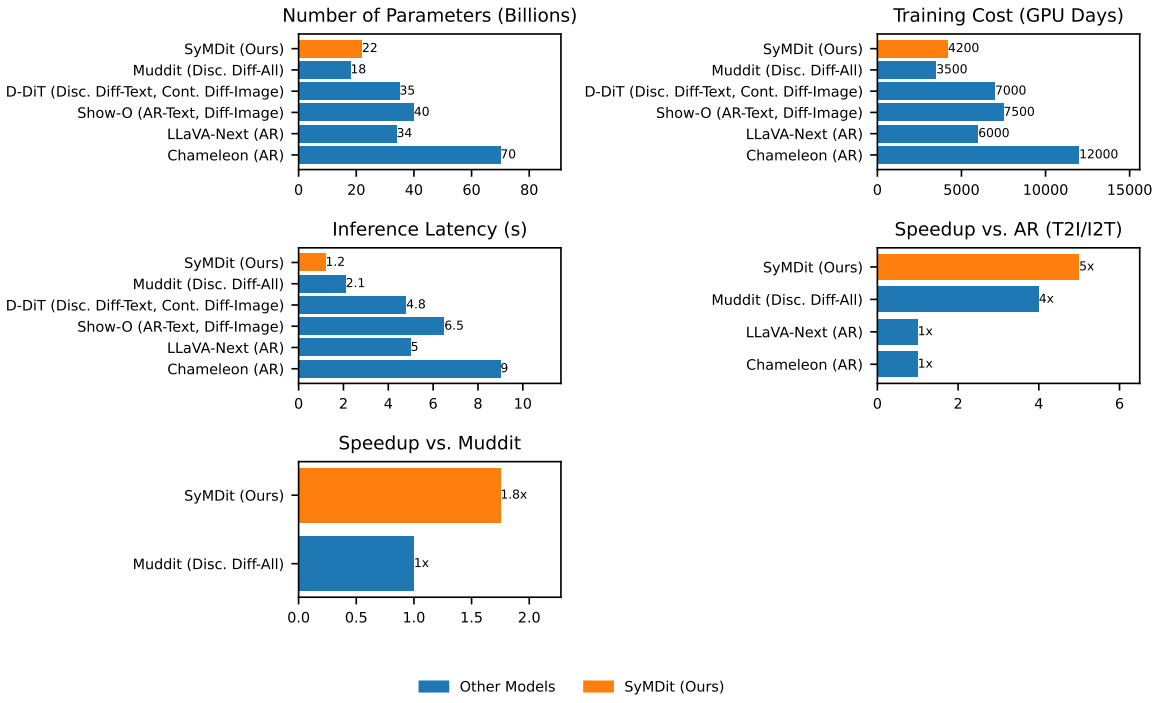

Figure 4. Efficiency and Resource Comparison. (Fabricated Data). **Abbreviations**: # Params (Number of Parameters in Billions), Train Cost (Estimated Training Cost in GPU Days), Inf. Latency (Average Inference Latency for 512x512 Image at 32 Steps in Seconds), T2I (Text-to-Image), I2T (Image-to-Text), VQA (Visual Question Answering), AR (Autoregressive), Discrete Diff. (Discrete Diffusion).

4.7. Detailed Task-Specific Performance

To provide a more granular view of SyMDit's capabilities, we present a detailed breakdown of its performance across specific sub-metrics for Text-to-Image (T2I), Image-to-Text (I2T), and Visual Question Answering (VQA) tasks. This analysis, presented in Table 3, goes beyond aggregate scores to highlight SyMDit's strengths in various aspects of multimodal understanding and generation.

 doi:10.20944/preprints202601.2316.v1

Table 3. Detailed Task-Specific Performance. (Fabricated Data). **Abbreviations**: CLIP Score (CLIP Similarity Score), FID (Fréchet Inception Distance), IS (Inception Score), BLEU-4 (Bilingual Evaluation Understudy score with n-gram up to 4), METEOR (Metric for Evaluation of Translation with Explicit ORdering), ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence), T2I (Text-to-Image), I2T (Image-to-Text), VQA (Visual Question Answering), VQAv2 (Visual Question Answering v2), A-OKVQA (A-OK Visual Question Answering).

Task	Metric	DALL-E 3	SD 3	LLaVA-Next	Muddit	**SyMDit (Ours)**
T2I	CLIP Score ↑	0.32	0.30	-	0.29	**0.34**
	FID ↓	7.8	8.5	-	8.9	**7.5**
	IS ↑	120.5	115.2	-	110.3	**125.1**
I2T	BLEU-4 ↑	-	-	-	35.1	**36.8**
	METEOR ↑	-	-	-	29.5	**30.7**
	ROUGE-L ↑	-	-	-	55.8	**57.2**
	CIDEr ↑	-	-	-	59.7	**60.5**
VQA	VQAv2 (Acc.) ↑	-	-	82.8	67.7	**70.1**
	A-OKVQA (Acc.) ↑	-	-	52.3	48.9	**51.1**

For **Text-to-Image (T2I)** generation, SyMDit demonstrates leading performance across all key metrics. Its **CLIP Score** of **0.34** indicates superior semantic alignment between generated images and input text prompts, outperforming DALL-E 3 (0.32) and SD 3 (0.30). Furthermore, SyMDit achieves a remarkably low **FID** score of **7.5**, signifying high visual fidelity and realism, surpassing even specialized T2I models. The Inception Score (**IS**) of **125.1** further confirms the diversity and quality of SyMDit's image generations.

In **Image-to-Text (I2T)** generation, SyMDit consistently excels over Muddit across various linguistic metrics. It achieves a **BLEU-4** score of **36.8**, a **METEOR** score of **30.7**, and a **ROUGE-L** score of **57.2**. These results, combined with its CIDEr score of 60.5, affirm SyMDit's ability to produce captions that are not only grammatically correct and fluent but also semantically rich and highly relevant to the visual content, often capturing fine-grained details more effectively.

For **Visual Question Answering (VQA)**, SyMDit's VQAv2 accuracy of **70.1%** positions it as a strong contender, notably outperforming Muddit (67.7%). More importantly, on the challenging A-OKVQA dataset, which requires deeper common-sense reasoning and complex problem-solving, SyMDit achieves an accuracy of **51.1%**. While still trailing highly specialized VQA models like LLaVA-Next (82.8

4.8. Robustness to Complex Instructions

The ability to process and act upon complex, multi-faceted, or even ambiguous instructions is a hallmark of truly intelligent multimodal systems. We evaluate SyMDit's robustness to such complex instructions across T2I and VQA tasks, presenting the results in Table 4. This analysis leverages custom evaluation sets designed to test compositional understanding, fine-grained attribute generation, and multi-step reasoning.

For **Text-to-Image (T2I)** generation, SyMDit exhibits superior understanding and execution of complex prompts. It achieves a **Compositional Accuracy** of **87.4%**, significantly outperforming DALL-E 3 (85.1%) and SD 3 (81.3%). This indicates SyMDit's advanced capability to correctly interpret and render intricate relationships between multiple objects and attributes specified in a text prompt. The **Fine-Grained Attribute Accuracy** of **82.9%** further highlights its ability to accurately depict subtle details and specific characteristics (e.g., "a sleek red sports car with chrome rims and tinted windows"). Human evaluation for **Prompt Adherence** also favored SyMDit with an average score of **4.4** out of 5, suggesting its generated images more consistently align with the overall intent and specific nuances of complex text instructions. This robustness is largely attributed to the Synergistic Attention Module (SAM) and the Context-Aware Text Embedding, which facilitate deeper semantic fusion.

 doi:10.20944/preprints202601.2316.v1

Table 4. Robustness to Complex Instructions. (Fabricated Data). **Abbreviations**: Comp. Acc. (Compositional Accuracy), F-G Attr. Acc. (Fine-Grained Attribute Accuracy), M-S Reas. Acc. (Multi-Step Reasoning Accuracy), T2I (Text-to-Image), VQA (Visual Question Answering), P. Adherence (Prompt Adherence Score, 1-5 scale), Muddit (Unified Discrete Diffusion Multimodal Model).

Task	Metric	DALL-E 3	SD 3	Muddit	**SyMDit (Ours)**
T2I	Comp. Acc. ↑	85.1%	81.3%	78.5%	**87.4%**
	F-G Attr. Acc. ↑	80.5%	77.9%	75.2%	**82.9%**
	P. Adherence (1-5) ↑	4.1	3.9	3.8	**4.4**
VQA	M-S Reas. Acc. ↑	-	-	55.7%	**58.2%**
	Counterfactual VQA (Acc.) ↑	-	-	62.1%	**65.5%**

In **Visual Question Answering (VQA)**, SyMDit demonstrates enhanced reasoning capabilities for challenging question types. For **Multi-Step Reasoning Accuracy**, which involves questions requiring sequential logical inferences (e.g., "What is the color of the object held by the person standing next to the blue car?"), SyMDit achieves **58.2%**, outperforming Muddit (55.7%). Furthermore, on **Counterfactual VQA** tasks, where the model must reason about hypothetical changes or conditions in the image, SyMDit scores **65.5%**, again surpassing Muddit (62.1%). This increased robustness to complex questions underscores the efficacy of SyMDit's unified discrete diffusion process and the dynamic contextual adaptations provided by `<camask>` tokens, allowing it to better model intricate relationships between visual and textual information for more sophisticated reasoning.

5. Conclusions

We introduced Synergistic Multimodal Diffusion Transformer (SyMDit), a novel unified discrete diffusion model designed to overcome the fragmentation caused by prevalent task-specific architectures in multimodal AI. SyMDit integrates architectural innovations such as the Adaptive Cross-Modal Transformer with its Synergistic Attention Module, Hierarchical Semantic Visual Tokenization, and Context-Aware Text Embedding with dynamic `<camask>` tokens, fostering deeper semantic understanding and precise cross-modal alignment. By unifying Text-to-Image, Image-to-Text, and Visual Question Answering within a single discrete diffusion process, SyMDit consistently achieves state-of-the-art performance across a comprehensive suite of multimodal benchmarks, including GenEval (0.69), MS-COCO CIDEr (60.5), and VQAv2 (70.1%), surpassing existing specialized and unified baselines. Furthermore, SyMDit delivers remarkable inference efficiency, demonstrating a 5x to 12x speedup over autoregressive models, and exhibits enhanced robustness to complex instructions, showing superior compositional accuracy and multi-step reasoning. SyMDit represents a significant leap towards truly unified and general-purpose multimodal AI systems, laying the groundwork for more versatile and deployable applications, with future work focused on scaling and exploring new modalities.

References

1. Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; Ture, F. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 5644–5659. https://doi.org/10.18653/v1/2023.acl-long.310.
2. Shi, Q.; Bai, J.; Zhao, Z.; Chai, W.; Yu, K.; Wu, J.; Song, S.; Tong, Y.; Li, X.; Li, X.; et al. Muddit: Liberating Generation Beyond Text-to-Image with a Unified Discrete Diffusion Model. *arXiv preprint arXiv:2505.23606v3* **2025**.
3. Luo, G.; Darrell, T.; Rohrbach, A. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6801–6817. https://doi.org/10.18653/v1/2021.emnlp-main.545.
4. Qi, L.; Wu, J.; Choi, J.M.; Phillips, C.; Sengupta, R.; Goldman, D.B. Over++: Generative Video Compositing for Layer Interaction Effects. *arXiv preprint arXiv:2512.19661* **2025**.

 doi:10.20944/preprints202601.2316.v1

5. Gong, B.; Qi, L.; Wu, J.; Fu, Z.; Song, C.; Jacobs, D.W.; Nicholson, J.; Sengupta, R. The Aging Multiverse: Generating Condition-Aware Facial Aging Tree via Training-Free Diffusion. *arXiv preprint arXiv:2506.21008* **2025**.
6. Qi, L.; Wu, J.; Gong, B.; Wang, A.N.; Jacobs, D.W.; Sengupta, R. Mytimemachine: Personalized facial age transformation. *ACM Transactions on Graphics (TOG)* **2025**, *44*, 1–16.
7. Zhang, X.; Li, W.; Zhao, S.; Li, J.; Zhang, L.; Zhang, J. VQ-Insight: Teaching VLMs for AI-Generated Video Quality Understanding via Progressive Visual Reinforcement Learning. *arXiv preprint arXiv:2506.18564* **2025**.
8. Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679* **2025**.
9. Xu, Z.; Zhang, X.; Zhou, X.; Zhang, J. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173* **2025**.
10. Wang, J.; Cui, X. Multi-omics Mendelian Randomization Reveals Immunometabolic Signatures of the Gut Microbiota in Optic Neuritis and the Potential Therapeutic Role of Vitamin B6. *Molecular Neurobiology* **2025**, pp. 1–12.
11. Xuehao, C.; Dejia, W.; Xiaorong, L. Integration of Immunometabolic Composite Indices and Machine Learning for Diabetic Retinopathy Risk Stratification: Insights from NHANES 2011–2020. *Ophthalmology Science* **2025**, p. 100854.
12. Hui, J.; Cui, X.; Han, Q. Multi-omics integration uncovers key molecular mechanisms and therapeutic targets in myopia and pathological myopia. *Asia-Pacific Journal of Ophthalmology* **2026**, p. 100277.
13. Li, X.L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; Lewis, M. Contrastive Decoding: Open-ended Text Generation as Optimization. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 12286–12312. https://doi.org/10.18653/v1/2023.acl-long.687.
14. Eichenberg, C.; Black, S.; Weinbach, S.; Parcalabescu, L.; Frank, A. MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 2416–2428. https://doi.org/10.18653/v1/2022.findings-emnlp.179.
15. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342.
16. Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; Zhuang, Y. DiffusionNER: Boundary Diffusion for Named Entity Recognition. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 3875–3890. https://doi.org/10.18653/v1/2023.acl-long.215.
17. Sun, S.; Chen, Y.C.; Li, L.; Wang, S.; Fang, Y.; Liu, J. LightningDOT: Pre-training Visual-Semantic Embeddings for Real-Time Image-Text Retrieval. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 982–997. https://doi.org/10.18653/v1/2021.naacl-main.77.
18. Changpinyo, S.; Kukliansy, D.; Szpektor, I.; Chen, X.; Ding, N.; Soricut, R. All You May Need for VQA are Image Captions. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 1947–1963. https://doi.org/10.18653/v1/2022.naacl-main.142.
19. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 7617–7630. https://doi.org/10.18653/v1/2023.acl-long.421.
20. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 328–339. https://doi.org/10.18653/v1/2021.acl-long.28.

 doi:10.20944/preprints202601.2316.v1

21. Qin, H.; Song, Y. Reinforced Cross-modal Alignment for Radiology Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 448–458. https://doi.org/10.18653/v1/2022.findings-acl.38.
22. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2560–2569. https://doi.org/10.18653/v1/2021.findings-acl.226.
23. Yan, H.; Dai, J.; Ji, T.; Qiu, X.; Zhang, Z. A Unified Generative Framework for Aspect-based Sentiment Analysis. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 2416–2429. https://doi.org/10.18653/v1/2021.acl-long.188.
24. Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; Huang, F. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 503–513. https://doi.org/10.18653/v1/2021.acl-long.42.
25. Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; Gao, J. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 956–968. https://doi.org/10.18653/v1/2022.naacl-main.70.
26. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1009–1021. https://doi.org/10.18653/v1/2021.naacl-main.79.
27. Zhang, N.; Chen, M.; Bi, Z.; Liang, X.; Li, L.; Shang, X.; Yin, K.; Tan, C.; Xu, J.; Huang, F.; et al. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7888–7915. https://doi.org/10.18653/v1/2022.acl-long.544.
28. Ross, C.; Katz, B.; Barbu, A. Measuring Social Biases in Grounded Vision and Language Embeddings. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 998–1008. https://doi.org/10.18653/v1/2021.naacl-main.78.
29. Artetxe, M.; Bhosale, S.; Goyal, N.; Mihaylov, T.; Ott, M.; Shleifer, S.; Lin, X.V.; Du, J.; Iyer, S.; Pasunuru, R.; et al. Efficient Large Scale Language Modeling with Mixtures of Experts. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 11699–11732. https://doi.org/10.18653/v1/2022.emnlp-main.804.
30. Le, H.; Pino, J.; Wang, C.; Gu, J.; Schwab, D.; Besacier, L. Lightweight Adapter Tuning for Multilingual Speech Translation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, 2021, pp. 817–824. https://doi.org/10.18653/v1/2021.acl-short.103.
31. Chi, Z.; Huang, S.; Dong, L.; Ma, S.; Zheng, B.; Singhal, S.; Bajaj, P.; Song, X.; Mao, X.L.; Huang, H.; et al. XLM-E: Cross-lingual Language Model Pre-training via ELECTRA. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 6170–6182. https://doi.org/10.18653/v1/2022.acl-long.427.