Bridging the Discrete-Continuous Gap: Unified Multimodal Generation via Coupled Manifold Discrete Absorbing Diffusion

Anonymous ACL submission

Abstract

The bifurcation of generative modeling into autoregressive approaches for discrete data (text) and diffusion approaches for continuous data (images) hinders the development of truly unified multimodal systems. While Masked Language Models (MLMs) offer efficient bidirectional context, they traditionally lack the generative fidelity of autoregressive models and the semantic continuity of diffusion models. Furthermore, extending masked generation to multimodal settings introduces severe alignment challenges and training instability. In this work, we propose **CoM-DAD** (**Co**upled **M**anifold **D**iscrete **A**bsorbing **D**iffusion), a novel probabilistic framework that reformulates multimodal generation as a hierarchical dual-process. CoM-DAD decouples high-level semantic planning from low-level token synthesis. First, we model the semantic manifold via a continuous latent diffusion process; second, we treat token generation as a discrete absorbing diffusion process, regulated by a **Variable-Rate Noise Schedule**, conditioned on these evolving semantic priors. Crucially, we introduce a **Stochastic Mixed-Modal Transport** strategy that aligns disparate modalities without requiring heavy contrastive dual-encoders. Our method demonstrates superior stability over standard masked modeling, establishing a new paradigm for scalable, unified text-image generation.

1 Introduction

The pursuit of Artificial General Intelligence (AGI) necessitates models capable of reasoning and generating across diverse modalities. However, a fundamental topological disconnect persists in current architectures: language is inherently discrete and symbolic, while visual data is continuous and dense. Consequently, the field has fragmented into two dominant paradigms: Autoregressive (AR) models, which excel at discrete text generation, and Continuous Diffusion Models (CDMs), which dominate high-fidelity image synthesis.

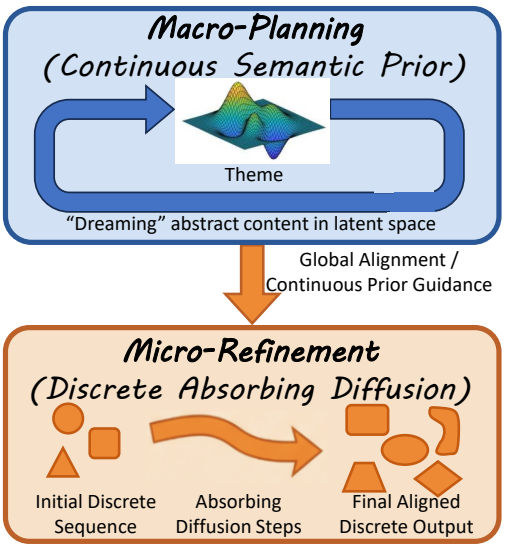


Figure 1: **Overview of CoM-DAD.** The framework splits generation into **Macro-Planning** (Top), where a continuous latent diffusion models abstract semantic "themes" (the "dreaming" phase), and **Micro-Refinement** (Bottom), where a discrete absorbing diffusion synthesizes tokens. The vertical arrow signifies the conditioning of the discrete generation process on the continuous prior, ensuring global alignment between the abstract plan and the final token sequence.

Efforts to unify these paradigms often result in compromised hybrids. Masked Generative Models (MGMs) attempt to bridge this gap by treating generation as a parallel denoising task. While MGMs offer significant efficiency gains over AR models (which suffer from serial dependency) and faster inference than CDMs, they notoriously struggle with **generative consistency**. Without a strong prior, masking-based models often produce locally coherent but globally disjoint outputs. Recent advancements, such as Representation-Conditioned Generation (RCG) (Li et al., 2024), have mitigated this in the visual domain by conditioning pixel generation on self-supervised representations. However, RCG remains strictly unimodal, focusing on

unconditional image synthesis and failing to address the complexities of cross-modal alignment or the discrete nature of language.

We argue that the difficulty in training multimodal masked models stems from forcing a single network to simultaneously learn semantic abstraction (what to generate) and structural composition (how to arrange tokens). To resolve this, we introduce CoM-DAD, a hierarchical framework that mathematically formalizes masked generation not as simple "filling in the blanks," but as a Discrete Absorbing Diffusion Process guided by a Continuous Semantic Prior. As conceptually illustrated in Figure 1, this hierarchical decoupling allows us to manage semantic abstraction and structural composition on separate, optimized manifolds. Our approach operates on two levels:

1. Macro-Planning (Latent Space): We employ a lightweight diffusion model to navigate the continuous manifold of high-level semantic representations. This allows the model to "dream" the abstract content of an image or sentence before committing to specific tokens.

2. Micro-Refinement (Discrete Space): We formulate token generation as a reverse diffusion process where tokens emerge from an absorbing state ([MASK]). Unlike standard MLMs, our transition kernel is strictly conditioned on the macro-plan, ensuring that every generated token is globally aligned with the intended semantic target. Crucially, this process is governed by a Variable-Rate Noise Schedule, which replaces fixed masking ratios with a continuous time parameter. This allows the model to learn generation from pure noise, while the transition kernel remains strictly conditioned on the macro-plan to ensure global alignment.

Furthermore, to address the scarcity of aligned multimodal data, we propose a **Stochastic Mixed-Modal Transport** mechanism during training. By dynamically swapping semantic priors between modalities (e.g., forcing the model to generate image tokens from a text representation), we induce a unified semantic space without the need for auxiliary alignment losses like CLIP.

Our contributions can be summarized as follows:

- We propose **CoM-DAD**, a unified probabilistic framework that bridges topological gaps between modalities by coupling a continuous latent diffusion for semantic planning with a discrete absorbing diffusion for synthesis.

- We introduce a **Variable-Rate Discrete Diffusion** mechanism that generalizes masked modeling with a continuous noise schedule, significantly improving generative consistency over static masking strategies.

- We develop a **Stochastic Mixed-Modal Transport** strategy that naturally aligns visual and textual manifolds via dynamic representation swapping, eliminating the need for heavy contrastive dual-encoder pre-training.

- We demonstrate that our hierarchical decoupling achieves superior training stability and sampling efficiency compared to standard autoregressive or monolithic diffusion baselines in multimodal contexts.

2 Related Work

Diffusion Models for Discrete Generation. Diffusion models have achieved notable success in continuous domains such as image and audio generation (Sohl-Dickstein et al., 2015; Ho et al., 2020). Extending diffusion to discrete sequences has attracted increasing interest (Li et al., 2025), particularly for language and symbolic reasoning. Early works (Hoogeboom et al., 2021; Austin et al., 2021) adapt diffusion to discrete spaces via relaxation or masking strategies. Later approaches (Li et al., 2022; He et al., 2022) embed discrete tokens into continuous spaces to enable Gaussian diffusion. Although effective, this introduces optimization difficulties, as simultaneously optimizing both the embedding layer and the diffusion model can lead to shortcut learning. In contrast, **CoM-DAD** operates directly on the discrete manifold. By introducing a hierarchical coupling with a continuous latent planner, it enables stable, efficient, and controllable generation without the optimization instability associated with joint embedding-diffusion training.

Masked Language Models as Discrete Absorbing Processes. Masked language modeling (MLM) has recently been reinterpreted as a form of discrete diffusion with absorbing states (Google DeepMind, 2025; Nie et al., 2025; Wu et al., 2025; Ye et al., 2025). While these methods effectively approximate autoregressive generation through iterative unmasking, they typically restrict both se-

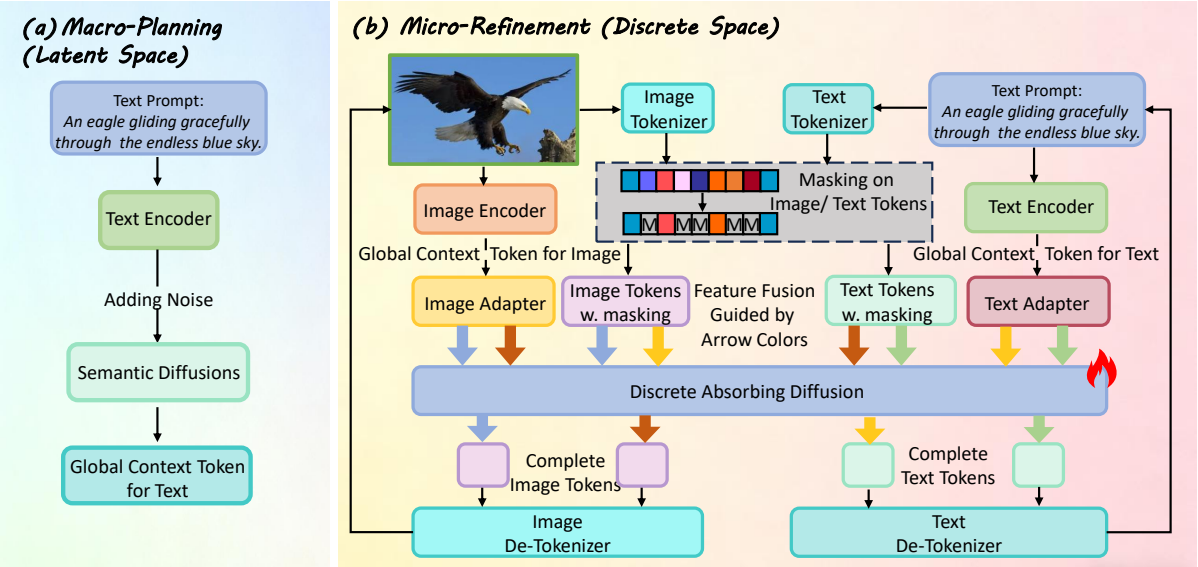


Figure 2: **The CoM-DAD Training Pipeline.** The framework consists of two coupled diffusion processes. **Left (Stage I):** The *Manifold-Constrained Semantic Diffusion* learns a continuous prior over semantic representations (r) via an SDE, capable of handling both text and image modalities. **Right (Stage II):** The *Semantic-Aware Discrete Absorbing Diffusion* reconstructs the discrete token sequence (x) from a masked state ($\tilde{x}_t$). The **Semantic Injection Interface** (center) connects these topologies by projecting the sampled semantic plan r into the decoder's embedding space, conditioning the reverse diffusion step $p_\theta(x_{t-1}|\tilde{x}_t, r)$ to ensure global semantic coherence. Cross-Modal Alignment is applied to (b).

mantic abstraction and structural composition to the discrete token space. **CoM-DAD** distinguishes itself by hierarchically decoupling these tasks. We employ a continuous latent diffusion to generate a high-level semantic plan, which serves as a robust condition for the discrete absorbing process, effectively bridging the topological gap between continuous semantic planning and discrete structural generation.

Multimodal Generation and Semantic Guidance. Multimodal generation aims to produce coherent outputs across heterogeneous data types. Prior work often relies on joint embedding spaces or autoregressive multimodal transformers (Alayrac et al., 2022; Chen et al., 2023; Team et al., 2023) to bridge the modality gap. **CoM-DAD** advances this paradigm by introducing Stochastic Mixed-Modal Transport. Rather than treating modalities as disparate sources to be aligned, we unify them into a shared continuous semantic manifold. This allows for dynamic prior swapping, where the discrete diffusion process is universally guided by the continuous planner, facilitating seamless cross-modal generalization. Semantic-Aware generation (Li et al., 2024; Wang and Torr, 2022) demonstrates that high-level representations can effectively guide continuous diffusion, though existing methods remain unimodal. **CoM-DAD** extends this idea to multimodal discrete generation by injecting learned representations directly into the absorbing diffusion process, enabling efficient semantic planning and cross-modal alignment within a unified framework.

3 Method

In this section, we formally detail **CoM-DAD**, a hierarchical generative framework that bridges the gap between continuous semantic exploration and discrete token generation. Unlike prior semantic-conditioned approaches such as RCG (Li et al., 2024), which focus exclusively on unimodal image synthesis using standard backbones, CoM-DAD introduces a **unified discrete diffusion mechanism** capable of joint text-image modeling.

Our framework, schematically illustrated in Figure 2, decomposes the intractable multimodal distribution $p(x)$ into two tractable generative processes operating in distinct topological spaces: (1) a **Continuous Latent Diffusion** process modeling the high-level semantic manifold $\mathcal{R}$, and (2) a **Discrete Absorbing Diffusion** process modeling the token-space conditional distribution $p(x|r)$. The figure highlights how the Semantic Injection Interface acts as the critical bridge, translating the continuous "plan" from the latent diffusion into a guiding signal for the discrete token denoiser.

3.1 Theoretical Formulation

Let $x \in \mathcal{X}$ represent a discrete sequence (e.g., text tokens or quantized image patches) and $r \in \mathbb{R}^d$

Table 1: **Quantitative comparison of unconditional text generation performance. CoM-DAD** outperforms existing autoregressive and masked language model baselines in both BLEU-2 and BLEU-4 metrics. This superior fidelity validates the effectiveness of the Continuous Latent Planner in maintaining global coherence across the discrete text manifold. "Ours (large) + Autoregressive" denotes a variant where CoM-DAD is constrained to generate tokens in a sequential order.

Methods	Pretained	Training Steps	Inference Iterations	Output Length	BLEU /% (↑)	Self-BLEU /% (↓)
Autoregressive Prediction						
GPT-2	✓	1M	40	40	10.81	40.02
BERT (base) (Devlin, 2018)	✓	1M	40	40	7.80	10.06
BERT (large) (Devlin, 2018)	✓	1M	40	40	5.05	**9.43**
Ours (base) + Autoregressive	×	400K+300K	20	256	8.52	18.37
Ours (large) + Autoregressive	×	400K+300K	20	256	**13.64**	16.52
Diffusion-based Methods						
D3PM (Austin et al., 2021)	×	1M	128	128	42.41	22.88
Diffusion-LM (Li et al., 2022)	×	740K	2000	64	35.53	26.68
DiffusionBERT (He et al., 2022)	×	1.9M	128	128	43.58	21.51
BERT-Mouth (Wang and Cho, 2019)	✓	2.8M	10	50	28.67	**12.4**
Ours (base)	×	400K+300K	20	256	29.42	18.37
Ours (large)	×	400K+300K	20	256	**47.46**	16.52

be a continuous semantic vector derived from a pre-trained encoder $\mathcal{E}(x)$. We maximize the evidence lower bound (ELBO) of the log-likelihood $\log p_\theta(x)$:

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}q(r|x)[\log p\theta(x|r)]}_{\text{reconstruction}} - \underbrace{D\text{KL}(q(r|x)|p_\phi(r))}_{\text{prior matching}}. \tag{1}$$

Unlike standard VAEs where the prior $p(r)$ is a static Gaussian, we parameterize $p_\phi(r)$ as a **continuous diffusion model**. Furthermore, we model the reconstruction term $p_\theta(x|r)$ not as a simple autoregressive decoder, but as a **discrete diffusion process** over the vocabulary set V. This hybrid formulation allows CoM-DAD to decouple global semantic planning (in continuous space $\mathcal{R}$) from local structural refinement (in discrete space $\mathcal{X}$).

3.2 Stage I: Manifold-Constrained Semantic Diffusion

The first stage models the prior distribution of semantic representations $p_\phi(r)$. While RCG (Li et al., 2024) utilizes a standard diffusion model for image class embeddings, we employ a modality-agnostic diffusion process capable of navigating the joint semantic space of both vision and language.

Given a semantic vector $r_0 = \mathcal{E}(x)$, we define a forward stochastic differential equation (SDE) that gradually destroys semantic information:

$$dr_t = -\frac{1}{2}\beta(t)r_t dt + \sqrt{\beta(t)}d\mathbf{w}_t, \tag{2}$$

where $\mathbf{w}_t$ is standard Brownian motion. We train a time-dependent denoiser $\epsilon_\phi(r_t, t)$ to reverse this process. Crucially, to ensure stability across modalities with varying norms, we employ **representation normalization** before the diffusion process. The training objective is the standard reweighted variational bound:

$$\mathcal{L}_{\text{latent}} = \mathbb{E}_{t,r_0,\epsilon}\left[\|\epsilon - \epsilon_\phi(r_t, t, c_m)\|^2\right], \tag{3}$$

where c_m indicates the modality ID, allowing the single model to learn the distinct manifolds of textual (r_{txt}) and visual (r_{img}) semantics simultaneously.

3.3 Stage II: Semantic-Aware Discrete Absorbing Diffusion

The core innovation of CoM-DAD is the formulation of sequence generation as a **Discrete Absorbing Diffusion Process**, generalizing the concept of "masked language modeling" into a rigorous probabilistic framework.

Discrete Forward Process. Let $x_0 = (w_1, \ldots, w_L)$ be a sequence of tokens from vocabulary V. We define a forward transition matrix Q_t that transitions any token w to a special absorbing state `[MASK]` with probability γ_t, and leaves it unchanged with probability $1 - \gamma_t$. This defines a corrupted sequence $\tilde{x}_t$ where a subset of tokens are masked according to the Markov property of the absorbing state.

Variable-Rate Noise Schedule. A critical component of our approach is the noise schedule $\gamma(t)$. Unlike the fixed masking strategies used in standard BERT-like models (typically 15%), we sample the masking ratio γ_t from a continuous time

schedule $t \sim U[0,1]$. This creates a **Variable-Rate Noise Schedule** that forces the model to learn generation across the entire complexity spectrum—from pure noise ($\gamma_1 \approx 1$) to fine-grained refinement ($\gamma_0 \approx 0$). This dynamic schedule is what effectively shifts the model's capability from simple local infilling to robust global generation.

Semantic-Aware Denoising. We learn a reverse transition kernel $p_\theta(x_{t-1}|\tilde{x}_t, r)$ parameterized by a Transformer. To condition the discrete generation on the continuous semantic vector r sampled from Stage I, we introduce a **Semantic Injection Interface**:

$$h_0 = [\text{Proj}(r); \text{Embed}(\tilde{x}_t)]. \tag{4}$$

By projecting r into the token embedding space and prepending it as a global context token, the Transformer attention mechanism allows every discrete decoding step to attend to the global semantic plan. The learning objective is the negative log-likelihood over the masked regions $\mathcal{M}$:

$$\mathcal{L}_{\text{discrete}} = \mathbb{E}_{x,t,r}\left[-\sum_{i\in\mathcal{M}} \log p_\theta(x_i|\tilde{x}_{\setminus\mathcal{M}}, r)\right]. \tag{5}$$

This formulation fundamentally differs from RCG, which relies on pixel-space diffusion or latent VAE decoders. By operating in discrete token space, it natively handles text and quantized images (via VQ-VAE tokens) in a unified architecture.

3.4 Cross-Modal Alignment via Inter-Modal Optimal Transport

A critical challenge in multimodal generation is the misalignment between visual and textual representation spaces. We address this via a **Mixed-Modal Sampling Strategy** that effectively approximates an optimal transport plan between modalities.

We construct training batches $\mathcal{B} = \{\mathcal{B}_{txt}, \mathcal{B}_{img}, \mathcal{B}_{pair}\}$.

- **Intra-Modal Learning:** For $\mathcal{B}_{txt}$ and $\mathcal{B}_{img}$, we train the model to reconstruct x given its own representation $r = \mathcal{E}(x)$.
- **Cross-Modal Bridge:** For paired data $\mathcal{B}_{pair}$, we perform **representation swapping**. We train the model to generate image tokens x_{img} conditioned on text representations r_{txt}, and vice-versa.

To facilitate this, we introduce lightweight **Modality Adapters** $\mathcal{A}_{T\to V}$ and $\mathcal{A}_{V\to T}$ that project representations into a shared semantic centroid before injection. This forces the latent diffusion model (Stage I) and the discrete generator (Stage II) to agree on a unified semantic coordinate system, enabling zero-shot generation (e.g., text-to-image) even with limited paired data.

4 Experiments

We empirically evaluate the effectiveness of **CoM-DAD** on both unimodal and cross-modal generative tasks. Our goals are threefold: (1) to assess the generative quality and efficiency of the discrete absorbing process on high-dimensional image and text manifolds, (2) to validate the hierarchical decoupling hypothesis by analyzing the impact of the Continuous Latent Planner and Semantic Injection Interface on convergence and stability, and (3) to investigate the efficacy of Stochastic Mixed-Modal Transport for zero-shot cross-modal alignment and generalization.

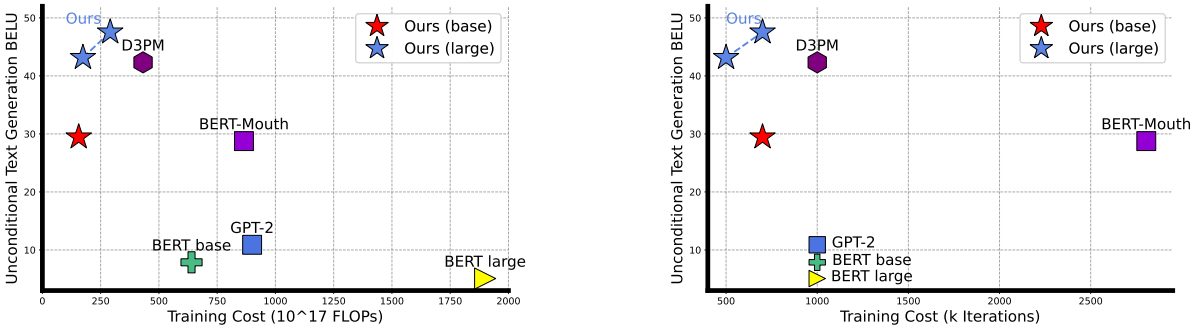
4.1 Implementation Details

Data Sources and Mixed-Modal Transport. Following our **Stochastic Mixed-Modal Transport** strategy (Sec. 3.4), we construct a unified training distribution comprising three subsets: (i) large-scale textual data from BookCorpus and Wikipedia (Wettig et al., 2022) (28M samples), (ii) image-only data from ImageNet-1k (1.28M samples), and (iii) 100K image-text pairs curated from COCO, with synthetic captions generated using Chameleon (Team, 2024). To ensure balanced manifold coverage and stable prior swapping, the sampling ratio across these sources is fixed at 2:2:1.

Discrete Manifold Tokenization. To establish the discrete state space, images are resized and center-cropped to 256×256 pixels, then tokenized using VQGAN (Yu et al., 2021) into 256 discrete visual tokens. Text inputs are tokenized using the RoBERTa tokenizer (Liu, 2019) and padded or truncated to a maximum of 256 tokens. No additional data augmentation is applied, relying on the variable-rate noise schedule for robustness.

Continuous Manifold and Optimization. We utilize frozen self-supervised encoders to define the continuous semantic manifold: MoCoV3 (Chen et al., 2021) for images and MPNet for text. The framework is trained in two decoupled stages:

- **Stage I (Continuous Latent Diffusion):** The continuous planner is trained 400K iterations to model the density of semantic embeddings.



(a) Training FLOPs versus BLEU performance.

(b) Convergence behavior across training iterations.

Figure 3: **Impact of the Semantic Injection Interface on convergence efficiency. CoM-DAD** achieves superior BLEU scores with substantially reduced training costs compared to the ablated variant without the interface. The reported cost includes training of both the *Continuous Latent Planner* (Stage I) and *Discrete Absorbing Diffusion* (Stage II). These results indicate that the **Semantic Injection Interface** accelerates convergence by enabling the discrete model to exploit the structure of the continuous semantic manifold, rather than learning it from scratch.

- **Stage II (Discrete Absorbing Diffusion):** The discrete generator is trained 300K iterations to generate tokens from absorbing states.

We use the AdamW optimizer with an initial learning rate of 5×10^{-4}. All experiments are conducted on 8 NVIDIA A800 GPUs.

Evaluation Protocol. We evaluate the topological unification capabilities of CoM-DAD. For unconditional image generation, we follow (Li et al., 2024) and generate 50K samples from the ImageNet distribution, reporting Inception Score (IS) (Salimans et al., 2016) and Fréchet Inception Distance (FID) (Heusel et al., 2017). For text generation, we report BLEU (n-gram accuracy) (Papineni et al., 2002) and Self-BLEU (diversity) (Zhu et al., 2018). For cross-modal generation, we test the Modality Adapters by prompting with text to synthesize aligned images, evaluating semantic consistency through qualitative visual inspection.

4.2 Main Results and Analysis

Superior fidelity on discrete text manifolds. We first evaluate **CoM-DAD** on unconditional text generation and compare it with strong baselines, including autoregressive models and standard masked language models. Table 1 shows that CoM-DAD achieves BLEU-2 and BLEU-4 scores of 47.46 and 13.64, respectively, outperforming all prior approaches under comparable settings. We also include a variant labeled "Ours (large) + Autoregressive", where CoM-DAD operates in a sequential manner, demonstrating the framework's flexibility to encompass autoregressive generation as a special case. Compared to standard autoregressive models such as GPT-2, CoM-DAD demonstrates exceptional stability in generating long-range dependencies from scratch, validating the efficacy of decoupling semantic planning from token generation. The generated text exhibits not only local fluency but also superior global coherence, a direct result of the Continuous Latent Planner governing the generation trajectory.

Semantic Injection Interface facilitates convergence. Figure 3 analyzes the relationship between generation quality and training cost. CoM-DAD achieves superior convergence rates with significantly fewer iterations than an ablated variant lacking the **Semantic Injection Interface**. This indicates that conditioning the discrete absorbing process on the continuous semantic manifold allows the model to bypass the difficulty of learning structure from scratch. Furthermore, models utilizing this interface produce longer and semantically richer sequences, whereas unguided counterparts tend to suffer from mode collapse or repetition, failing to bridge the gap between the discrete and continuous states effectively.

Parallel decoding via Discrete Absorbing Diffusion. Unlike autoregressive models that are bound by serial token decoding, CoM-DAD leverages a Discrete Absorbing Diffusion process that allows for non-autoregressive parallel generation. As shown in Figure 4(a), the model can reconstruct up to 20 tokens per step simultaneously, significantly reducing inference latency. Despite the absence of aggressive GPU-specific optimizations found in mature autoregressive systems, CoM-DAD achieves a $5\times$ speedup over GPT-2. This efficiency confirms that modeling generation as an iterative denoising process from absorbing states is a viable and high-throughput alternative to standard causal modeling.

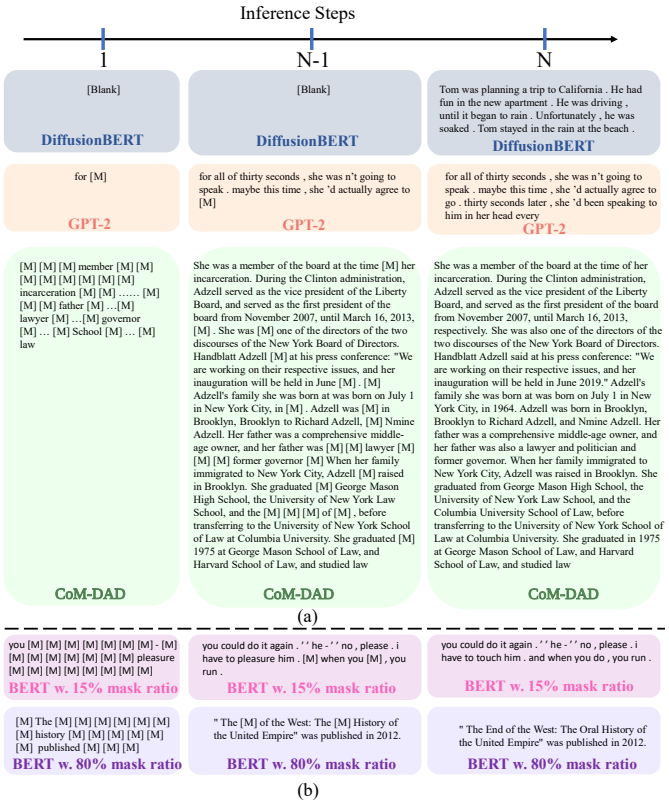


Figure 4: **Analysis of parallel decoding dynamics and schedule impact.** (a) Comparison of generation paradigms: **CoM-DAD** utilizes *Discrete Absorbing Diffusion* for efficient non-autoregressive parallel decoding, contrasting with the serial nature of autoregressive baselines. (b) Ablation on absorbing rates: Models trained with the aggressive **Variable-Rate Noise Schedule** (High Masking) demonstrate emergent semantic prioritization (*main-first, details-later*), establishing global structure before local details. Zoomed-in views are provided for clarity.

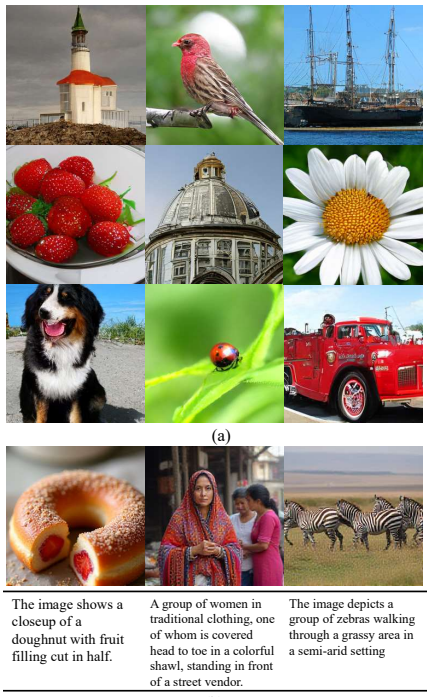


Figure 5: **Visualization of D-MLLM's image generation capabilities.** (a) Unconditional image generation results demonstrating diversity and visual quality. (b) Text-to-image generation showcasing effective cross-modal alignment, with synthesized images accurately reflecting the semantic content of input text prompts.

Emergent semantic prioritization via Variable-Rate Noise Schedule. An interesting emergent behavior observed in CoM-DAD is its preference for resolving high-information content in early diffusion steps, followed by lower-entropy details—a pattern we term *main-first, details-later*. In Figure 4(a), the Variable-Rate Discrete Noise Schedule forces the model to first anchor key subject-verb structures before filling in modifiers and syntactic connectors. This validates our hypothesis that the noise schedule dictates the hierarchy of generation.

To isolate this effect, we compare models trained with different absorbing rates (Figure 4(b)). Only models trained with an aggressive noise schedule inherent to CoM-DAD develop this prioritized behavior, producing more globally consistent outputs. This confirms that the variable-rate schedule does not merely add noise, but actively encourages the model to learn a hierarchical data decomposition.

Topological unification and Mixed-Modal Transport. We evaluate CoM-DAD on unconditional and text-driven image generation to assess its cross-modal capabilities. As shown in Table 2, CoM-DAD achieves competitive Inception Scores (IS) and Fréchet Inception Distances (FID), specifically demonstrating that the Stochastic Mixed-Modal Transport strategy successfully aligns heterogeneous manifolds without degradation.

Figure 5 visualizes outputs from CoM-DAD. Unconditional samples (Figure 5(a)) are visually coherent, while text-driven samples (Figure 5(b)) exhibit strong semantic alignment with the prompts. These results demonstrate that our unified continuous semantic manifold allows for effective Dynamic Prior Swapping, enabling the discrete image generator to be accurately guided by textual plans without requiring massive paired datasets.

4.3 Ablation Studies

To further understand the role of each component in the **CoM-DAD** framework, we conduct a series of ablations focusing on the hierarchical decoupling

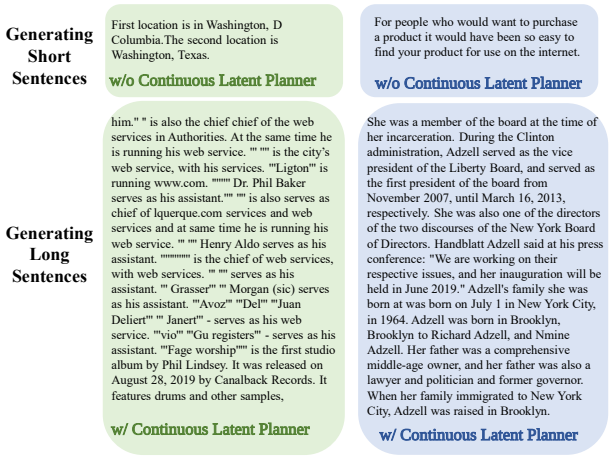
Table 2: **Quantitative comparison of unconditional image generation. CoM-DAD** achieves competitive performance against state-of-the-art baselines on the continuous image manifold. These results validate the framework's *Topological Unification*, demonstrating that the *Stochastic Mixed-Modal Transport* strategy effectively aligns discrete token generation with continuous visual semantics.

Unconditional Generation	params	FID (↓)	IS (↑)
BigGAN (Brock et al., 2018)	70M	38.61	24.7
IC-GAN (Casanova et al., 2021)	75M	15.6	59
ADM (Dhariwal and Nichol, 2021)	554M	26.21	39.7
ADDP (Tian et al., 2023)	176M	8.9	95.3
MaskGIT (Chang et al., 2022)	227M	20.72	42.1
RDM-IN (Blattmann et al., 2022)	400M	5.91	158.8
MAGE-B (Li et al., 2023)	176M	8.67	94.8
MAGE-L (Li et al., 2023)	439M	7.04	123.5
RCG-B (Li et al., 2024)	239M	3.98	177.8
RCG-L (Li et al., 2024)	502M	3.44	186.9
Ours (base)	318M	5.14	138.3
Ours (large)	593M	4.32	151.6

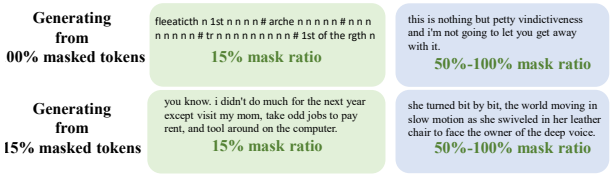
and the diffusion noise schedules.

Impact of the Continuous Latent Planner. To validate the necessity of topological decoupling, we remove the *Continuous Latent Planner* (Stage I) and train the *Discrete Absorbing Diffusion* model directly on token sequences without the *Semantic Injection Interface*. As illustrated in Figure 6a, models trained without continuous latent conditioning can produce short, syntactically correct phrases but often fail when tasked with generating longer or semantically complex passages. Furthermore, they require significantly more training iterations to achieve comparable quality. This supports our fundamental hypothesis that externalizing semantic planning into a continuous manifold simplifies the discrete learning objective, allowing the token generator to focus solely on mapping high-level plans to discrete structures.

Necessity of the Variable-Rate Noise Schedule. We retrain CoM-DAD using a standard fixed 15% masking rate (typical of BERT-style MLMs). As shown in Figure 6b, this variant fails to generate coherent text from the fully absorbing state, instead repeating simple or semantically trivial tokens. In contrast, the **Variable-Rate Noise Schedule** employed in CoM-DAD successfully generates complete sentences from scratch. This confirms that aggressive, variable-rate masking is essential for shifting the model from a local infilling objective to a global generative task. These findings are consistent with our insight in Sec. 3.3 and provide empirical support for *Discrete Absorbing Diffusion* as a mechanism for robust generation rather than mere masked prediction.



(a) Ablation analysis of CoM-DAD components.



(b) Impact of Noise Schedules.

Figure 6: **Ablation studies of CoM-DAD.** (a) Removing the **Continuous Latent Planner** (and the corresponding **Semantic Injection Interface**) substantially degrades the model's ability to generate complex, long-form passages, leading to slower convergence and reduced global coherence. This supports our hypothesis that topological decoupling simplifies the learning objective for the **Discrete Absorbing Diffusion** process. (b) Comparison of the **Variable-Rate Noise Schedule** against a fixed 15% rate. The fixed-rate variant yields repetitive, trivial outputs when starting from a *fully absorbing state*. In contrast, CoM-DAD's variable schedule enables generation from scratch, proving that dynamic masking is critical to shift from local infilling to global generation.

5 Conclusion

Summary of Benefits. CoM-DAD's architectural and training design confers several benefits: (1) **Efficiency:** By decoupling representation modeling from token generation, CoM-DAD enables faster convergence and 5× faster inference over standard denoising or autoregressive models. (2) **Generative Capability:** The variable masking schedule fosters the model's ability to generate coherent and diverse outputs rather than merely recover corrupted inputs. (3)**Multimodal Alignment:** Our mixed sampling strategy facilitates scalable training from unimodal data while achieving strong cross-modal consistency. (4)**Unified Architecture:** A single encoder-decoder model handles both text and image generation through a shared conditioning mechanism, supporting flexible and generalizable generation tasks.

Limitations

While CoM-DAD effectively bridges the topological gap between discrete and continuous modalities, our current empirical validation is primarily focused on foundational image-text generation tasks. Although the Stochastic Mixed-Modal Transport framework is theoretically extensible to temporal modalities like video or audio, we reserve the specific calibration of the Variable-Rate Noise Schedule for these high-dimensional domains for future work to maintain focused analysis. Furthermore, we observe that standard automated metrics may not fully capture the long-horizon semantic consistency driven by the Continuous Latent Planner, potentially underrepresenting the model's ability to generate conceptually accurate but structurally diverse outputs compared to rigid token-matching baselines.

References

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

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. 2021. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and 1 others. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649.

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

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

Google DeepMind. 2025. Gemini diffusion. https://deepmind.google/models/gemini-diffusion. Accessed: 2025-05-24.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2022. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. 2023. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152.

Tianhong Li, Dina Katabi, and Kaiming He. 2024. Return of unconditional generation: A self-supervised representation generation method. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Xiao Li, Jiaqi Zhang, Shuxiang Zhang, Tianshui Chen, Liang Lin, and Guangrun Wang. 2025. In-situ tweedie discrete diffusion models. *arXiv preprint arXiv:2510.01047*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Changyao Tian, Chenxin Tao, Jifeng Dai, Hao Li, Ziheng Li, Lewei Lu, Xiaogang Wang, Hongsheng Li, Gao Huang, and Xizhou Zhu. 2023. Addp: Learning general representations for image recognition and generation with alternating denoising diffusion process. *arXiv preprint arXiv:2306.05423*.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Guangrun Wang and Philip HS Torr. 2022. Traditional classification neural networks are good generators: They are competitive with ddpms and gans. *arXiv preprint arXiv:2211.14794*.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b.

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

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.