Masked Diffusion Captioning for Visual Feature Learning

Chao Feng[1,2] **Zihao Wei**[1,3] **Andrew Owens**[1,2]

[1]University of Michigan [2]Cornell University [3]University of Maryland

https://cfeng16.github.io/mdlm4vfl/

cf583@cornell.edu

Abstract

We learn visual features by captioning images with an image-conditioned masked diffusion language model, a formulation we call masked diffusion captioning (MDC). During training, text tokens in each image–caption pair are masked at a randomly chosen ratio, and a decoder conditioned on visual features is trained to reconstruct the original text. After training, the learned visual features can be applied to downstream vision tasks. Unlike autoregressive captioning, the strength of the visual learning signal in MDC does not depend on each token's position in the sequence, reducing the need for auxiliary objectives. Linear probing experiments across a variety of academic-scale models and datasets show that the learned visual features are competitive with those produced by autoregressive and contrastive approaches.

1 Introduction

Multimodal models that learn the cross-modal associations between images and language have driven many recent advances in visual representation learning (Desai and Johnson, 2021; Radford et al., 2021; Tschannen et al., 2023). An intuitively appealing approach is to pose this problem as *visual captioning*: first, train an image-conditioned language model to generate text captions from images, and then use its learned visual features for downstream tasks. However, the popular formulation of captioning as autoregressive language modeling often yields visual features that perform worse than those from alternative vision-language learning approaches. One major reason for this is the asymmetry in the learning signal (Tschannen et al., 2023): later text tokens can be predicted so well from the earlier ones that the image becomes decreasingly important as the sequence progresses from left to right. A variety of approaches have addressed this issue by augmenting the objective with right-to-left generation (Desai and Johnson, 2021),

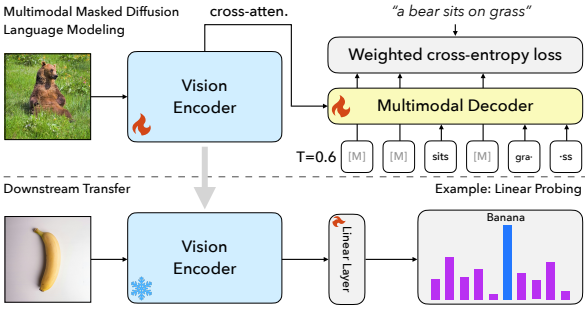


Figure 1: **Learning visual features by masked diffusion language modeling.** We learn visual features by captioning images using an image-conditioned masked diffusion language model. After training, features from the visual encoder can be transferred to downstream computer vision tasks.

contrastive learning (Yu et al., 2022), and parallel decoding (Tschannen et al., 2023) objectives.

An emerging line of work in the natural language processing community has applied masked diffusion language models (MDLMs) to text generation (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024). Instead of producing text in a fixed order, these methods randomly mask tokens at each iteration and train a model to reconstruct the original text. During training, the fraction of masked tokens is chosen randomly, enabling the model to reconstruct text given arbitrary numbers of masked tokens. Previous work has shown that such models can generate high-quality text via ancestral sampling, optimize variational bounds, and learn language features that transfer well to downstream tasks (Sahoo et al., 2024).

In this paper, we learn visual features through *masked diffusion captioning* (MDC): using an image-conditioned masked diffusion language model to generate text captions from images (Fig. 1). Unlike autoregressive models, the amount of text conditioning each token receives is not determined by its position in the sequence; instead, each token provides a position-independent amount of visual supervision. Since we primarily use captioning as a means of learning features rather than

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25247–25263
November 4-9, 2025

as an end in itself, our approach is closely related to methods that learn visual features with image-conditioned BERT (Sariyildiz et al., 2020). However, instead of using a fixed masking ratio, we sample ratios randomly during training and weight the loss as a function of the ratio.

We evaluate our approach on academic-scale models and datasets, establishing an effective training recipe for masked diffusion captioning. Our experiments suggest that the resulting model learns useful visual features across multiple datasets and encoder architectures (e.g., CC12M (Changpinyo et al., 2021) with ViT-B and ViT-L (Dosovitskiy et al., 2020)). These features achieve performance that is competitive with autoregressive and contrastive methods on a variety of linear probing experiments for visual recognition tasks. We also find that the model's ability to approximately estimate the likelihood of a given caption can be used to match images to their captions successfully, resulting in competitive performance on compositionality benchmarks (Hsieh et al., 2023; Yuksekgonul et al., 2022). Additionally, we find that image-conditioned BERT, a special case of our model, can achieve features competitive with those of other learning approaches when properly tuned, typically by choosing a large masking ratio that requires the model to rely heavily on the visual signal.

2 Related Work

Image captioning for visual representation learning. Contrastive vision-language pretraining (Radford et al., 2021; Tschannen et al., 2025; Zhai et al., 2023; Yu et al., 2022; Sun et al., 2023; Bolya et al., 2025) learns strong visual features through the discriminative task of contrastive learning. There is a line of work that seeks to obtain good visual representations by captioning, where the model is supervised at the token level. This paradigm of feature learning through generative pretraining can produce both visual features and captioning models capable of generating text for specific tasks. VirTex (Desai and Johnson, 2021) utilizes forward (left-to-right) and backward (right-to-left) captioning to learn visual features. SimVLM (Wang et al., 2021) treats visual patches as the prefix and employs a single prefix language modeling objective for supervision. BLIP (Li et al., 2022a) uses synthetic captions to improve the quality of image-text pairs. Similarly, CoCa (Yu et al., 2022) leverages both contrastive learning and image captioning objectives. Recently, CapPa (Tschannen et al., 2023) has shown that captioning can produce strong visual encoders as competitive as those from contrastive learning on large datasets. It augments the autoregressive captioning objective with parallel decoding (i.e., where all tokens are masked, and the model must reconstruct the text). Following this direction, LocCa (Wan et al., 2024) and SigLIP 2 (Tschannen et al., 2025) employ captioning as a pretraining task. Additionally, there is prior work (Li et al., 2022a; Lai et al., 2024; Li et al., 2024b; Fan et al., 2023; Chen et al., 2024a; Singla et al., 2024) that aims to improve text quality for image-text pairs. Like these approaches, we learn visual features via image captioning, but we do so using a single masked diffusion language modeling objective, instead of an autoregressive or hybrid approach.

Vision language models. Contrastive learning methods, such as CLIP (Radford et al., 2021), have provided scalable and effective approaches for image-language learning. Large-scale datasets (Schuhmann et al., 2022; Gadre et al., 2023; Ordonez et al., 2011; Changpinyo et al., 2021; Sharma et al., 2018; Krishna et al., 2017) have contributed significantly to this success. These models (Radford et al., 2021; Tschannen et al., 2025; Zhai et al., 2023; Yu et al., 2022; Sun et al., 2023; Bolya et al., 2025) can perform visual recognition (Antol et al., 2015; Russakovsky et al., 2015; Lin et al., 2014) in a zero-shot manner by computing similarities between image and text embeddings. Recently, with the advancement of large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023a; Bai et al., 2023; Liu et al., 2024a; Team et al., 2023, 2024), multimodal models (202, 2023; Wang et al., 2022; Liu et al., 2023; Hurst et al., 2024; Liu et al., 2024b; Bai et al., 2025; Chen et al., 2024b; Tong et al., 2024; Li et al., 2024a; Yang et al., 2023) have been developed that perform vision tasks through language, given visual input processed by vision encoders (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025). These models demonstrate a variety of multimodal capabilities (Yue et al., 2024; Liu et al., 2024c; Yu et al., 2024a; Masry et al., 2022; Gurari et al., 2018; Yu et al., 2024b; Hao et al., 2025; Li et al., 2025). Despite their success, they often fail to capture complex relationships between images and language, such as compositionality (Hsieh et al., 2023).

Autoregressive language models. Autoregressive language models factorize the joint probability of a sequence into a product of conditional next-token probabilities and are trained with maximum-likelihood estimation (teacher forcing) to predict each token given its left context. The paradigms of next-token prediction and GPT-style models (Radford et al., 2018, 2019; Brown et al., 2020) laid the foundation for the success of large language models (Achiam et al., 2023; Touvron et al., 2023a; Bai et al., 2023; Liu et al., 2024a; Team et al., 2023, 2024; Touvron et al., 2023b; Grattafiori et al., 2024; Guo et al., 2025). Using autoregressive models for image captioning is also common practice (Vinyals et al., 2015).

Diffusion language models. Diffusion models were first proposed by Sohl-Dickstein et al. (2015) and later popularized for continuous data by DDPM (Ho et al., 2020) and score matching (Song et al., 2020; Song and Ermon, 2019). More recently, diffusion-based language models have gained significant attention. These methods can be broadly divided into two categories: (1) embedding-space diffusion (Li et al., 2022b) and (2) discrete-state diffusion (He et al., 2022; Austin et al., 2021; Hoogeboom et al., 2021; Lou et al., 2023; Sahoo et al., 2024; Shi et al., 2024; Zheng et al., 2024; Ou et al., 2024; Nie et al., 2024, 2025). Sohl-Dickstein et al. (2015) first introduced diffusion models with discrete state spaces over binary random variables, which were extended by Hoogeboom et al. (2021) to categorical data using uniform categorical noise. D3PM (Austin et al., 2021) introduced various transition matrices (uniform, absorbing, discretized Gaussian, and token embedding distance) for discrete-time Markov chains, while Campbell et al. (2022) extended this to continuous-time Markov chains (CTMC). Concrete score matching (Meng et al., 2022) generalized score matching (Song and Ermon, 2019) to discrete domains, and SEDD (Lou et al., 2023) further proposed score entropy for optimization. Both MDLM (Sahoo et al., 2024) and MD4 (Shi et al., 2024) derived simplified expressions of the ELBO for masked diffusion language models. Other work (Zheng et al., 2024; Ou et al., 2024) has suggested that input time embeddings are unnecessary for discrete diffusion language models. More recently, SMDM (Nie et al., 2024) demonstrated the scalability of masked diffusion language models, and LLaDA (Nie et al., 2025) scaled them to relatively large sizes. Building on these advances, our paper focuses on applying masked diffusion language models to visual representation learning through image captioning.

Vision-language masked modeling. A variety of recent methods have learned visual feature learning used masked language modeling (Li et al., 2019; Sun et al., 2019; Tan and Bansal, 2019; Li et al., 2020b,a; Lu et al., 2019; Chen et al., 2020; Su et al., 2019; Zhou et al., 2020; Li et al., 2021). However, these methods have typically focused on learning joint visual-linguistic representations through early fusion. Sariyildiz et al. (2020) first identifies candidate tokens in the caption that correspond to visual concepts, typically nouns, adjectives, or verbs, then randomly masks one of them and trains the model to predict it using both the image and the surrounding text. Similarly, Geng et al. (2022) and Swerdlow et al. (2025) extend masked modeling to both vision and language. In contrast, our work mainly focuses on learning visual representations from the captioning objective only by using masked diffusion language modeling. It avoids the need to choose a single (possibly dataset-dependent) masking ratio, and can directly generate text.

3 Method

We propose to learn visual features by generating text captions from images using an image-conditioned masked diffusion language model, an approach we call masked diffusion captioning (MDC).

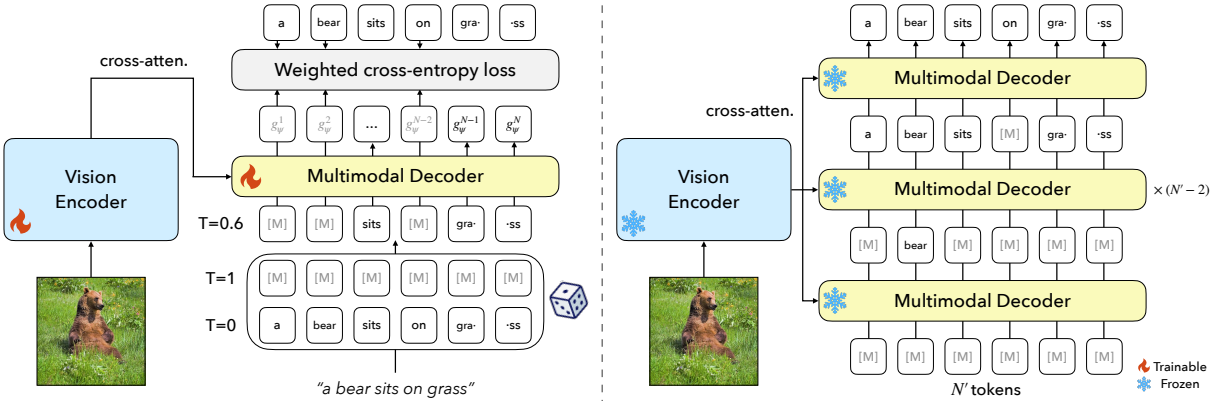
3.1 Preliminaries

We review masked diffusion language modeling.

Masked language modeling. The popular Bidirectional Transformer (BERT) (Devlin et al., 2019; Liu et al., 2019) model learns language representations via masked language modeling (MLM). Given a sequence $x^{1:N}$, a mask set M of token indices is sampled and forms a corrupted sequence $\tilde{x}^{1:N}$ by replacing tokens in M with `[MASK]` (or a random/unchanged token). The training loss is:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|M|} \sum_{i \in M} \log p_\theta \left(x^i \middle| \tilde{x}^{1:N} \right). \quad (1)$$

Masked diffusion language model (MDLM). MDLM (Sahoo et al., 2024) converts BERT-style models into generative masked diffusion models. Let x_0 be a text token with K categories, where



(a) Masked diffusion captioning (b) Image-conditioned language sampling

Figure 2: **Learning visual features using masked diffusion captioning.** (a) We train an image-conditioned masked diffusion language model to learn visual features. Given an image and its corresponding text caption, we randomly mask text tokens in the caption. We then reconstruct the caption, using a decoder that is conditioned on visual features (obtained from a separate encoder network) and the text tokens. In each training iteration, we sample a time step t that determines a masking ratio and a cross-entropy weight. $T = 0$ means no masked token while $T = 1$ means sequence is fully masked. (b) During sampling, we start with a fully masked sequence containing N' mask tokens. We then iteratively denoise N' steps to obtain a full caption.

K is the size of the vocabulary $\mathcal{X} = 1, \ldots, K$ ($K = |\mathcal{X}|$). MDLM adds a `[MASK]` token to the vocabulary (as an absorbing state), which functions similarly to the mask token used in BERT (Devlin et al., 2019) and in conditional masked language models (Ghazvininejad et al., 2019).

For time steps r and t with $r < t$, the forward process is:

$$q(\mathbf{x}_t|\mathbf{x}_r) = \begin{cases} \delta_{x_t,\texttt{[MASK]}}, & \text{if } \mathbf{x}_r = \mathbf{m} \\ \text{Cat}\left(\mathbf{x}_t; \frac{\alpha_t}{\alpha_r}\mathbf{x}_0 + \left(1 - \frac{\alpha_t}{\alpha_r}\right)\mathbf{m}\right), & \text{if } \mathbf{x}_r \neq \mathbf{m} \end{cases} \quad (2)$$

where δ is the delta function, $\mathbf{x}_t$ is the one-hot encoding of x_t on timestep t, $\mathbf{m}$ is the one-hot encoding of `[MASK]`, and α_t is the predefined noise schedule between 0 and 1, which is a strictly decreasing function of t. At each time step, $\mathbf{x}_t$ remains unchanged or transitions to `[MASK]`, determined by transition probability. The posterior can be expressed as:

$$q(\mathbf{x}_r|\mathbf{x}_t, \mathbf{x}_0) = \begin{cases} \text{Cat}\left(\mathbf{x}_r; \frac{(1-\alpha_r)\mathbf{m} + (\alpha_r - \alpha_t)\mathbf{x}_0}{1-\alpha_t}\right), & \text{if } \mathbf{x}_t = \mathbf{m} \\ \delta_{x_r,x_t}, & \text{if } \mathbf{x}_t \neq \mathbf{m}. \end{cases} \quad (3)$$

We train the language model μ_θ to reconstruct masked tokens given the unmasked ones. The training objective (Sahoo et al., 2024) computes the weighted cross entropy loss for each masked token. The per-token loss can be written as:

$$\mathcal{L}_{\text{NELBO}} = \mathbb{E}_t\left[\frac{\alpha_t'}{1-\alpha_t}\mathbb{E}_q\left[\delta_{x_t,\texttt{[MASK]}}\mathbf{x}_0^\top \log\left(\mu_\theta^i\left(\mathbf{x}_t^{1:N}, t\right)\right)\right]\right] \quad (4)$$

where $\mathbf{x}_0$ is the one-hot encoding for the token (i.e., the ground truth for the reconstruction).

3.2 Masked diffusion captioning

We apply the masked diffusion language modeling to the problem of visual captioning, with the goal of learning visual features.

Training. Each training pair consists of image $I \in \mathbb{R}^{3\times H\times W}$ and its corresponding caption $C = [c^0, \ldots, c^{N-1}]$. We use a standard transformer encoder-decoder architecture following Tschannen et al. (2023) as the captioner h. Encoder f_ϕ takes image I and produces a sequence of visual features $\mathbf{V} = f_\phi(I) = [\mathbf{v}_0, \ldots, \mathbf{v}_{M-1}]$. These are (late) fused with the decoder g_ψ by cross attention to predict caption C.

Building on the training objective of the MDLM (Sahoo et al., 2024), we define the loss for our masked diffusion captioning (MDC). Given the caption C, MDC chooses a factorized forward process $q\left(C_t|C_0, \mathbf{V}\right) = \prod_{i=0}^{N-1} q\left(c_t^i|c_0^i, \mathbf{V}\right)$, the learned reverse process is also factorized $p_\psi\left(C_r|C_t, \mathbf{V}\right) := \prod_{i=0}^{N-1} q\left(c_r^i|c_t^i, g_\psi^i\left(C_t, t, \mathbf{V}\right)\right)$. Thus, the training objective is:

$$\mathcal{L}_{\text{MDC}} = \mathbb{E}_t\left[\frac{\alpha_t'}{1-\alpha_t}\mathbb{E}_q\left[\sum_{i=0}^{N-1}\delta_{c_t^i,\texttt{[MASK]}}\mathbf{c}_0^{i\top}\log\left(g_\psi^i\left(C_t, t, \mathbf{V}\right)\right)\right]\right], \quad (5)$$

Following recent work (Zheng et al., 2024; Sahoo et al., 2024; Ou et al., 2024; Nie et al., 2025, 2024) we adopt a time-independent model parameterization. We omit t from the input for text decoder g_ψ, while the entire captioner h still uses the noise as part of the loss weight $\frac{\alpha_t'}{1-\alpha_t}$ (Eq. 5). We use a linear schedule (Lou et al., 2023; Sahoo et al., 2024; Shi et al., 2024) for α_t, where $\alpha_t = 1 - t$. The training process is also presented in Alg. 1.

Algorithm 1 Pseudocode of training for masked diffusion captioner.

```
# imgs: batch of images
# caps: batch of corresponding captions
# img_enc: vision encoder in captioner
# text_dec: text decoder in captioner
# t: sampled time step in [0,1] for noise schedule
# B: batch size of minibatch
# L: sequence length for minibatch
# MASK: mask token ID
for imgs, caps in loader: # load a minibatch
    img_feats = image_enc(imgs) # sequence of visual tokens
    t = uniform(B, 1)
    p = uniform(B, L)
    masked_caps = caps.clone()
    masked_caps[p < t] = MASK
    logits = text_dec(masked_caps, img_feats)
    loss = (1/t) * cross_entropy(logits[p < t], caps[p < t])
    loss.backward()
```

Sampling. Once the captioner h is trained, we can not only use its visual encoder f_ϕ for downstream tasks but also the decoder g_ψ to generate text. Beyond this, we can also use the variational lower bound $\log p_\psi(C|f_\phi(I))$ to perform classification tasks, by comparing the probability of different captions (Sec. 4.2). It has been revealed that there are numerical instability issues in Gumbel-based categorical sampling (Zheng et al., 2024), so we choose to use the token-by-token decoding strategy inspired by (Ghazvininejad et al., 2019; Chang et al., 2022; Nie et al., 2025; Zheng et al., 2024; Nie et al., 2024) for image captioning. Specifically, with a predefined sequence length of N' generated tokens, masked diffusion captioning employs N' denoising steps. Starting from a fully masked sequence, the denoiser (decoder) performs predictions for all masked tokens at each iteration.

We use greedy decoding for our captioning experiments. At each masked position, we use the maximum probability assigned by the model across its vocabulary as a proxy for the confidence score of the predicted token. In each iteration, the single masked token with the overall highest confidence score across all predictions is then revealed (i.e., unmasked). All other tokens that were masked at the beginning of the iteration remain masked for the subsequent iteration:

$$x_{t-1}^i = \begin{cases} x_t^i, & \text{if } x_t^i \neq \texttt{[MASK]} \\ \text{argmax}\left(\eta^i\right), & \text{if } \max_j \eta_j^i > \max_{y \neq i}(\max_k \eta_k^y) \\ \texttt{[MASK]}, & \text{otherwise} \end{cases} \quad (6)$$

where $\eta_j^i = g_\psi^i(\mathbf{x}_t^{1:N'}, f_\phi(I))_j$.

Once a token is unmasked, it remains fixed throughout the rest of the denoising process. This strategy can make sure all [MASK] tokens are unmasked at the end of the denoising process. Compared with Gumbel-based categorical sampling, this denoising strategy is more efficient since no intermediate denoising step is wasted. Additionally, the denoising process can refine generated captions and mitigate uncertainties in parallel decoding. The training and sampling processes are also illustrated in Fig. 2.

4 Experiments

We pretrain our models and benchmark them against other approaches.

4.1 Implementation Details

Pretraining data. We pretrain models on three vision–language datasets: Conceptual 3M (CC3M) (Sharma et al., 2018), Conceptual 12M (CC12M) (Changpinyo et al., 2021), and subsets of Recap-DataComp (Li et al., 2024b). Because of its relatively small scale, CC3M is used primarily for schedule search. Both CC3M and CC12M are directly scraped from the Internet, whereas Recap-DataComp (Li et al., 2024b) is constructed by recaptioning the original DataComp dataset (Gadre et al., 2023) with Llama 3 (Grattafiori et al., 2024). Figure 3 presents the caption length distributions across these datasets.

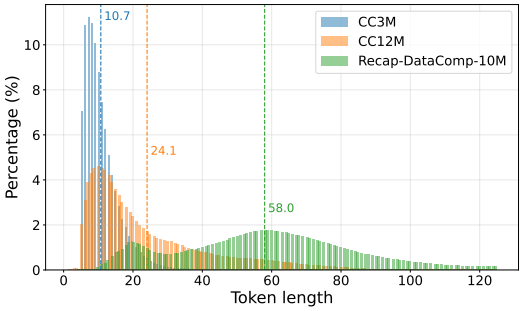


Figure 3: **Dataset caption length distribution.** We visualize caption length distribution for (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), and a 10M randomly sampled subset of Recap-DataComp (Li et al., 2024b) after tokenization.

Pretraining details and baselines. We pretrain three vision-language models from scratch for evaluation: CLIP (Radford et al., 2021), autoregressive captioning (ARC), and masked diffusion captioning (MDC), all implemented based on the OpenCLIP (Cherti et al., 2023). To ensure a fair comparison, all models are trained with the same set of hyperparameters.

For captioning models (ARC and MDC), we use ViT-B/32, ViT-B/16, and ViT-L/14 (Dosovitskiy et al., 2020) as the vision encoder backbones. For ViT-B, the multimodal text decoder is a 12-layer

Transformer decoder with 8 attention heads and a hidden size of 512, where each layer sequentially performs text self-attention, followed by image-text cross-attention. For ViT-L, multimodal text decoder consists of 12 layers with 12 attention heads and hidden size of 768. For text self-attention, ARC employs causal self-attention, while MDC utilizes bidirectional self-attention. Additionally, during training of MDC, only non-padding tokens are used as supervision signals, which can ensure the fair comparison between ARC and MDC. We use $[0.5, 1.0]$ as the default noise schedule of t for MDC. The CLIP models share the same vision backbones but replace the text decoders with Transformer encoders of the same architecture as multimodal text decoders. Input images are resized to 224×224, and text sequences are padded or truncated to 77 tokens.

We optimize all models using the AdamW optimizer (Loshchilov and Hutter, 2017) (see hyperparameter setups in the Appendix) and cosine learning rate decay. Training is conducted with a batch size of 128 per GPU (64 for ViT-L with 2 gradient accumulation steps). Specifically, we use 8 NVIDIA L40S GPUs for training.

4.2 Benchmarking masked diffusion captioning

Learning from image alt-text pairs. We first train all methods with ViT-B/32 and ViT-B/16 on CC12M (Sharma et al., 2018) and use linear probing to evaluate visual representations. Following prior work (Tschannen et al., 2023), we use global average pooling (GAP) of the encoder output sequence for visual representations to evaluate autoregressive captioning and masked diffusion captioning models. The feature of `[CLS]` token (pre-logits layer) is used for CLIP. We use CLIP-benchmark (LAION-AI, 2023) across standard datasets including ImageNet-1k (Russakovsky et al., 2015), Food101 (Bossard et al., 2014), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and Pets (Vedaldi, 2012). See hyperparameter setups for linear probing in the Appendix. As shown in Tab. 1, masked diffusion captioning achieves performance comparable to autoregressive captioning in terms of average accuracy, demonstrating that it can learn visual representations from image alt-text pairs effectively.

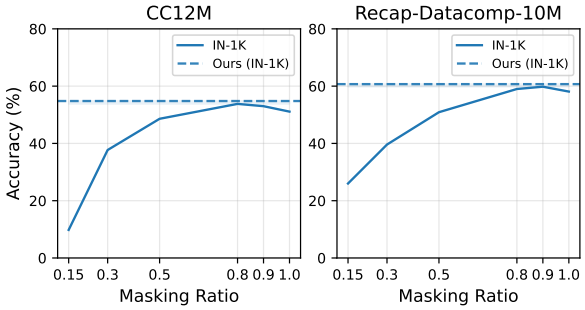


Figure 4: **Comparison to image-conditioned BERT with different masking ratios.** We compare our method against BERT with varying masking ratios, including 100% (parallel decoding). While BERT with certain masking ratios achieves performance close to ours, our method adopts a unified schedule, avoiding the need to tune the masking ratio on each dataset.

Learning from rich textual descriptions. Many captions in CC12M (Changpinyo et al., 2021) are noisy and not descriptive enough. To test the capability of models to learn from rich textual descriptions, we pretrain models on a randomly selected 10M subset of Recap-DataComp (Li et al., 2024b) mentioned in Sec. 4, where the length distribution is presented in Fig. 3. We use linear probing to evaluate learned features. As reported in Tab. 1, even when trained with twice the default batch size (denoted as CLIP-LB), CLIP struggles to learn strong visual features from detailed captions, consistent with prior findings (Li et al., 2024b; Zhang et al., 2024). Results in Tab. 1 suggest that masked diffusion captioning can learn effective visual features from descriptive captions. Surprisingly, scaling from ViT-B/16 to ViT-L/14 results in degraded performance for autoregressive captioning. Since ViT-L/14 incorporates a larger text decoder with stronger text modeling capacity, this may amplify the dependency issue (Tschannen et al., 2023), where later tokens rely more on previously generated tokens than on visual features. In contrast, performance of masked diffusion captioning boosts, indicating it has potential model size scalability for rich textual descriptions due to the random ordering.

Comparison with masked language model variants. We compare our masked diffusion captioning to other masked model variants: 1) BERT with varying masking ratios and 2) Parallel Decoding with 100% masking ratio. Results of linear probing are presented in Fig. 4. When masking ratio is low, such as 15%, the model can often reconstruct masked tokens using surrounding context, particularly when the masked ones are se-

Table 1: **Linear probing results.** To test the learned visual features, we evaluate CLIP, autoregressive captioning (ARC), and masked diffusion captioning (MDC) on several benchmarks by linear probing. Note that CLIP-LB uses twice the default batch size during pretraining, and its performance is shown in gray. The best results are in **bold**, and the second best are colored in blue. The evaluation metric is accuracy.

Backbone	Dataset	Method	ImageNet-1K	Food101	CIFAR-10	CIFAR-100	Pets	Average
ViT-B/32	CC12M	CLIP	**57.2**	66.4	**89.2**	**70.9**	**74.8**	**71.7**
		ARC	54.2	**67.7**	87.5	68.3	70.0	69.5
		MDC (Ours)	54.8	64.5	88.4	69.3	66.7	68.7
	Recap-DataComp-10M	CLIP-LB	55.5	66.4	91.0	75.4	66.1	70.9
		CLIP	53.1	66.0	90.5	75.0	63.9	69.7
		ARC	**61.4**	**76.0**	**94.1**	**79.1**	**70.7**	**76.3**
		MDC (Ours)	60.7	72.1	93.9	78.6	67.6	74.6
ViT-B/16	CC12M	CLIP	**67.3**	76.5	91.5	74.7	**82.3**	**78.5**
		ARC	64.7	**79.0**	91.1	72.8	77.0	76.9
		MDC (Ours)	65.9	76.0	**91.6**	**75.1**	77.0	77.1
	Recap-DataComp-10M	CLIP-LB	62.8	74.4	92.4	77.8	73.6	76.2
		CLIP	60.4	73.1	92.3	77.2	71.0	74.8
		ARC	**69.5**	**84.5**	**95.4**	81.3	72.4	**80.6**
		MDC (Ours)	69.0	81.3	95.2	**81.6**	**73.9**	80.2
ViT-L/14	Recap-DataComp-10M	CLIP-LB	64.8	76.3	93.4	78.6	75.4	77.7
		CLIP	62.1	75.1	93.1	78.1	73.2	76.3
		ARC	66.3	78.7	94.4	78.4	72.3	78.0
		MDC (Ours)	**71.6**	**83.4**	**95.3**	**81.4**	**83.8**	**83.1**

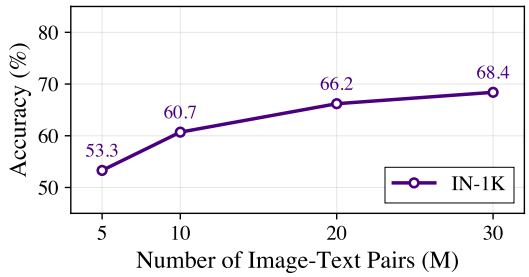


Figure 5: **Linear probing performance with varying numbers of image–text pairs.** We randomly sample 5M, 10M, 20M, and 30M pairs from Recap-DataComp-1B (Li et al., 2024b) for pretraining our method. As the number of image–text pairs increases, the linear probing performance on IN-1K improves.

mantically uninformative words like "a" or "the". This shortcut limits the model's reliance on visual input and hinders the learning of meaningful visual representations. In contrast, Parallel Decoding (100% masking ratio), which masks all tokens and requires them to be generated simultaneously based solely on the image, entirely ignores language structure. This not only impairs the model's ability to capture linguistic patterns but also burdens it with the dual challenge of learning both language structure and visual features. As a result, the pretraining task becomes more difficult, leading to slower convergence and weaker visual representations. Thus, by tuning the masking ratio for each dataset (Fig. 4), a high-ratio setting could be found that balances shortcut avoidance and task difficulty, yielding good performance. However, our method uses a unified time-based schedule that eliminates the need for such tuning. This design consistently outperforms fixed-ratio BERT variants and demonstrates the robustness of our masked diffusion captioning.

Dataset size scaling. We randomly sample 5M, 10M, 20M, and 30M image-text pairs from Recap-DataComp-1B (Li et al., 2024b) to pretrain our method with ViT-B/32 as the visual backbone. Linear probing results on IN-1k are shown in Fig. 5. We find that the more image text pairs used for pretraining, the better performance on the downstream tasks. This validates the potential dataset size scalability of our method.

Vision language compositionality. As mentioned in Sec. 3.2, captioning models can use their likelihood (Tschannen et al., 2023) or its variational bound to perform classification tasks in a zero-shot manner. Evaluating the compositionality of vision language models is a binary classification task. Given an image I, one correct caption $C_r = [c^{r_0}, \cdots, c^{r_{N_r-1}}]$, and one manipulated false caption $C_d = [c^{d_0}, \cdots, c^{d_{N_d-1}}]$, models need to recognize the true caption C_r. Autoregressive captioners can use factorization of joint probability as an indicator for binary classification:

$$\log\left(p_\psi\left(c^0, \cdots, c^{N-1}\right)\right) = \sum_{i=0}^{N-1} \log\left(p_\psi\left(c^i | c^{<i}, f_\phi(I)\right)\right), \tag{7}$$

f_ϕ is the visual encoder mentioned in Sec. 3.2. For masked diffusion captioner, we use lower

Table 2: **Vision language compositionality evaluation.** We evaluate compositionality of models on two benchmarks: ARO (Yuksekgonul et al., 2022) and SugarCrepe (Hsieh et al., 2023).

Method	ARO				SugarCrepe		
	relation	attribute	coco order	flickr30k order	add	replace	swap
CLIP	53.6	59.7	27.2	29.5	66.5	72.8	61.3
ARC	82.7	76.0	97.7	98.4	97.6	77.4	76.9
MDC (Monte Carlo)	**85.1**	**84.3**	89.0	89.0	85.6	75.8	75.2
MDC (Heuristic)	84.6	81.2	**98.4**	**98.8**	**97.8**	**77.9**	**78.5**

Table 3: **Image captioning evaluation.** We evaluate autoregressive captioning (ARC), masked diffusion captioning (MDC), and CoCa (Yu et al., 2022) on MSCOCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015) (B@4: BLEU@4 (Papineni et al., 2002), M: METEOR (Banerjee and Lavie, 2005), C: CIDEr (Vedantam et al., 2014), S: SPICE (Anderson et al., 2016), RL: ROUGE-L (Lin, 2004)). Performance of CoCa is represented in gray for reference. *While we report autoregressive captioning performance metrics, we note that the autoregressive model does not have access to the target sequence length during the generation process, in contrast to MDC, and as a result their performance is not directly comparable.

Method	sequence length		MSCOCO					Flickr30k				
	MSCOCO	Flickr30k	B@4	M	C	S	RL	B@4	M	C	S	RL
CoCa	–	–	21.95	21.41	67.61	20.92	43.12	–	–	–	–	–
ARC*	–	–	16.0	23.9	48.8	17.4	38.6	10.1	20.2	19.3	13.4	30.8
MDC	10	10	**20.7**	14.3	**64.7**	16.0	**42.2**	11.3	17.9	20.5	10.8	30.3
MDC	15	15	17.6	23.1	51.1	18.3	40.7	**15.3**	21.4	**28.6**	14.9	**35.0**
MDC	20	20	13.6	**31.9**	24.1	**18.6**	37.0	13.8	**21.7**	20.6	**15.5**	34.3

bound. For each caption C, $C_n = [c_n^0, \cdots, C_n^{N-1}]$ denotes C with n masked tokens. Therefore, the lower bound is (Ou et al., 2024; Nie et al., 2025; Zheng et al., 2024):

$$\log\left(p_\psi\left(c^0, \cdots, c^{N-1}\right)\right) \geq \sum_{n=1}^{N} \mathbb{E}_q\Big[\frac{1}{n}\sum_{i=0}^{N-1} \delta_{c_n^i,\texttt{[MASK]}} \log\left(p_\psi\left(c_0^i | C_n, f_\phi(I)\right)\right)\Big], \quad (8)$$

We use monte carlo estimate for t to get lower bound for true and false captions, where we set the number of forward processes to 1024 for each caption. Then the lower bound can be used as a proxy for classification. In addition, since our classification task requires a discriminative score rather than a full perplexity measure (which can be computationally demanding), we propose a more efficient heuristic variant that also achieves better performance. Starting with a fully masked sequence, we perform N denoising steps, equivalent to the caption length $|C|$. In each step, we identify the masked position with the highest predicted confidence (see Sec. 3.2) and record the log-likelihood of the ground-truth token from caption C at this position. This ground-truth token then replaces [MASK], and the updated sequence is fed into the subsequent step. The sum of these N log-likelihoods constitutes the final classification score. We evaluate all models on ARO (Yuksekgonul et al., 2022) and SugarCrepe (Hsieh et al., 2023) benchmarks. As presented in Tab. 2, MDC outperforms CLIP and ARC, suggesting that the masked diffusion training approach enhances vision-language compositionality in models.

Image captioning. To evaluate image captioning capability of autoregressive captioning and masked diffusion captioning, we finetune them on MSCOCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015) respectively, where they are both pretrained on CC12M (Changpinyo et al., 2021). For reference, we also test a publicly available pretrained and finetuned (on MSCOCO) checkpoint of CoCa (Yu et al., 2022) with the same vision backbone for reference. Due to the limitation of masked diffusion language models, vanilla masked diffusion captioning can only generate captions with a fixed sequence length, so we need to specify the output length at the beginning of sampling. Therefore, we use three variants of MDC: 10 tokens, 15 tokens, and 20 tokens for output. We use greedy decoding as mentioned in Sec. 3.2. Beam search with a beam size of 6 is employed for autoregressive captioning. Results in Tab. 3 demonstrate that masked diffusion captioning can sample reasonable captions. One possible reason it outperforms autoregressive captioning is that it uses more FLOPs. Moreover, the predefined sequence length of masked diffusion captioning might favor length-sensitive evaluation metrics. However, the comparison of captioning capability between autoregression and masked diffusion needs further research.

4.3 Analysis of Design Choices

We analyze certain design choices of masked diffusion captioning by linear probing on ImageNet-1k.

Necessity of t. To assess the necessity of t in the training objective, we perform an ablation study by removing t from the weighted cross-entropy loss during pretraining. The model then essentially becomes CMLM (Ghazvininejad et al., 2019). The results, presented in Tab. 4, show linear probing performance drops for models trained on both CC12M and Recap-DataComp-10M. This suggests that the loss scaling factor t plays a critical role in learning effective visual representations.

Table 4: **Ablation on t.** We compare masked diffusion captioning (MDC) with its loss variant pretrained on CC12M and Recap-DataComp-10M. We evaluate them by linear probing on ImageNet-1K.

Method	CC12M	Recap-DataComp-10M
MDC (w/o t)	54.0	59.5
MDC	**54.8**	**60.7**

Noise schedule. During the training of masked diffusion models, the noise level (masking ratio) of each step is determined by t sampled from the interval $[\omega_l, \omega_u]$. The vanilla masked diffusion model with linear schedule uses $\omega_l = 0, \omega_u = 1$. However, we find that loss is very unstable when pretrained on CC3M, where many captions are very short. This resonates with findings from prior work (Arriola et al., 2025). Thus, we conduct the ablation study on CC3M to analyze the effect of sampling interval of t, and the results are shown in Tab. 5. We find that $\omega_l = 0.5, \omega_u = 1$ achieves best performance and use this sampling interval as the default for masked diffusion captioning.

Table 5: **Analysis of noise schedule.** We test masked diffusion captioning pretrained on CC3M with different noise schedules by linear probing on ImageNet-1K.

Schedule	$[0.0, 1.0]$	$[0.3, 0.8]$	$[0.4, 0.9]$	$[0.5, 1.0]$
IN1k Acc.	29.3	36.1	38.6	**39.2**

5 Limitations

Both the pretraining dataset scale (on the order of 10M image-caption pairs) and the model size are at the academic scale. Training masked diffusion captioning on datasets that contain undesirable contents may result in the learning of biased or harmful visual representations and the generation of malicious captions.

6 Conclusion

In this work, we introduce masked diffusion captioning (MDC), an image-conditioned masked diffusion language model designed to learn visual representations. Our results demonstrate that masked diffusion captioning effectively learns visual features, outperforming previous masked language modeling variants by using a unified noise schedule. In addition, masked diffusion captioning can generate reasonable captions and exhibits strong vision-language compositionality. We conduct evaluations to establish an effective training recipe for masked diffusion captioning. Overall, our study suggests that masked diffusion language models offer a compelling alternative to autoregressive approaches for learning visual representations from image caption pairs.

Acknowledgements. This work was supported by Advanced Research Projects Agency for Health (ARPA-H) under award #1AY2AX000062. This research was funded, in part, by the U.S. Government. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We thank Subham Sahoo and Zixuan Pan for helpful discussions.

References

2023. Gpt-4v(ision) system card.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. *ArXiv*, abs/1607.08822.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*.

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, and 1 others. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. 2022. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279.

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.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, and Ramakrishna Vedantam. pycocoevalcap. https://github.com/salaniz/pycocoevalcap.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223.

Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the*

North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and 1 others. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112.

Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. 2022. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.

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*.

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.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Juan Lao Tebar, Wenze Hu, Zhe Gan, Peter Grasch, and 1 others. 2024. Revisit large-scale image-caption data in pre-training multimodal foundation models. *arXiv preprint arXiv:2410.02740*.

LAION-AI. 2023. Clip benchmark. https://github.com/LAION-AI/CLIP_benchmark.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, and 1 others. 2024a. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. 2025. Towards visual text grounding of multimodal large language model. *Preprint*, arXiv:2504.04974.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022b. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. 2024b. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pages 121–137. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2023. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. 2022. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35:34532–34545.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. 2024. Scaling up masked diffusion models on text. *arXiv preprint arXiv:2410.18514*.

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*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, ICVGIP '08, page 722–729, USA. IEEE Computer Society.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2024. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and 1 others. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning visual representations with caption annotations. In *European conference on computer vision*, pages 153–170. Springer.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*.

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.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *Preprint*, arXiv:1904.01766.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. 2025. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

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*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. 2023. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855.

Andrea Vedaldi. 2012. Cats and dogs. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, page 3498–3505, USA. IEEE Computer Society.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim M Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. 2024. Locca: Visual pretraining with location-aware captioners. *Advances in Neural Information Processing Systems*, 37:116355–116387.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024a. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. 2024b. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. 2024. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

A Hyperparameters for pretraining

Here we present the hyperparameters we used for pretraining the models with ViT-B and ViT-L based on datasets in Table 6, Table 7, and Table 8. All models share the same hyperparameter for pretraining. The exception is that we apply aggressive gradient norm clipping as shown in Table 8 during training of autoregressive captioning with ViT-L/14. Without this, the loss becomes numerically unstable. Additionally, we struggle to learn effective visual features for autoregressive captioning with ViT-L/14 on CC12M (Changpinyo et al., 2021) due to the same reason, where we need to use 0.1 for gradient norm clipping and 1e-4 for learning rate. All the results provided in the paper are just a single run with the given parameters.

Table 6: **Hyperparameters used to train vision-language models with ViT-B/32.**

config	CC12M	Recap-DataComp-10M
optimizer	AdamW	AdamW
base lr	5e-4	5e-4
warmup steps	10,000	10,000
weight decay	0.1	0.1
β_1	0.9	0.9
β_2	0.98	0.98
batch size	1024	1024
lr schedule	Cosine	Cosine
epochs	32	32

Table 7: **Hyperparameters used to train vision-language models with ViT-B/16.**

config	CC12M	Recap-DataComp-10M
optimizer	AdamW	AdamW
base lr	5e-4	5e-4
warmup steps	10,000	10,000
weight decay	0.2	0.2
β_1	0.9	0.9
β_2	0.98	0.98
batch size	1024	1024
lr schedule	Cosine	Cosine
epochs	32	32

Table 8: **Hyperparameters used to train vision-language models with ViT-L/14.**

config	Recap-DataComp-10M
optimizer	AdamW
base lr	4e-4
warmup steps	10,000
weight decay	0.2
β_1	0.9
β_2	0.98
batch size	1024
lr schedule	Cosine
epochs	32
grad norm (for AR only)	0.5

B Hyperparameters for Linear probing

The linear probing hyperparameters are set as the default values provided in CLIP-benchmark (LAION-AI, 2023), with a batch size of 64, 10 epochs, and a learning rate of 0.1.

C Diffusion preliminary

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020) have the forward and reverse Markov processes. Given a clean instance $\mathbf{x}_0$ (e.g., image) from the target distribution, forward process gradually corrupts it $\mathbf{x}_0\mathbf{x}_1 \ldots \mathbf{x}_T$ by $\mathbf{x}_t \sim q\left(\mathbf{x}_t|\, \mathbf{x}_{t-1}\right)$. For instance, Gaussian noise is gradually added: $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\,\mathbf{I})$. The learned reverse process $p_\theta\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right)$ can move the instance $\mathbf{x}_T$ sampled from source distribution towards target distribution. The training objective of variational lower bound for p_θ is:

$$\mathcal{L} = \mathbb{E}_q\Big[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\|p(\mathbf{x}_T))}_{L_T} + \sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\|p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Big] \quad (9)$$

D Qualitative result of captioning

We present some qualitative results in Fig. 6, revealing an interesting pattern: MDC can generate more descriptive words for captions when longer sampling lengths are specified.

E Scientific artifact

In this project, all the dataset we used and their license are in Tab. 9. We also adapted our training and evaluation code from OpenCLIP (Cherti et al., 2023) and CLIP-benchmark (LAION-AI, 2023). These codebases are under the MIT License.

Table 9: **Licenses for datasets used in this work.**

Dataset	License
ImageNet-1k (Russakovsky et al., 2015)	Custom (Non-commercial, research only)
STL-10 (Coates et al., 2011)	BSD License
Food101 (Bossard et al., 2014)	MIT License
VOC2007 (Everingham et al., 2010)	CC BY 4.0
CIFAR-10 (Krizhevsky et al., 2009)	MIT License
CIFAR-100 (Krizhevsky et al., 2009)	MIT License
Flowers (Nilsback and Zisserman, 2008)	CC BY 4.0
Pets (Vedaldi, 2012)	CC BY 4.0
MSCOCO (Lin et al., 2014)	CC BY 4.0
Flickr30k (Plummer et al., 2015)	Custom (Academic use only)
ARO (Yuksekgonul et al., 2022)	MIT License
SugarCrepe (Hsieh et al., 2023)	MIT License
CC3M (Sharma et al., 2018)	Custom (Use with attribution to Google LLC)
CC12M (Changpinyo et al., 2021)	Custom (Use with attribution to Google LLC)
Recap-DataComp-10M (Li et al., 2024b)	CC BY 4.0

Image			
Reference	A living area features an orange chair, a green chair, several wood chairs and a light colored sofa.	A train car sits idle on messy train tracks.	The view from an airplane seat displaying a bed of clouds.
CoCa	A living room filled with furniture and a flat screen tv mounted on the wall.	A blue and yellow train traveling down train tracks near a building.	The view of the wing of an airplane in the air.
ARC	A living room with a large brick wall and two green couches.	A blue and yellow passenger train sitting in a train yard.	The wing of an airplane as seen over a body of water.
MDC-10	A living room with a couch and television.	A blue train is on a train track.	An airplane wing high up in the sky.
MDC-15	A living room with a couch, a chair and a television.	A blue and yellow train traveling down tracks next to a building.	The wing of a plane is above the clouds in the sky.
MDC-20	A living room with a couch, a chair, a television and pictures on the wall.	A blue and yellow train with a blue engine on the tracks in front of a building.	A view from the wing of an airplane shows a view of the clouds in the sky.

Figure 6: **Examples of captioning results.** We show three examples sampled from MSCOCO Karpathy-test split. MDC-10/15/20 means the length of the output sequence is 10/15/20 for masked diffusion captioners.

F Packages

We use pycocoevalcap (Chen et al.) to evaluate image captioning.

G AI usage

We use ChatGPT for revising the grammar of the writing.