# UNIFIED DISCRETE DIFFUSION FOR SIMULTANEOUS VISION-LANGUAGE GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The recently developed discrete diffusion model performs extraordinarily well in generation tasks, especially in the text-to-image task, showing great potential for modeling multimodal signals. In this paper, we leverage these properties and present a unified multimodal generation model, which can perform text-based, image-based, and even vision-language simultaneous generation using a single model. Specifically, we unify the discrete diffusion process for multimodal signals by proposing a unified Markov transition matrix and a unified objective. Moreover, we design a multimodal mutual attention module to highlight the inter-modal linkages, which is vital for multimodal generation. Extensive experiments indicate that our proposed method can perform comparably to the state-of-the-art solutions in various generation tasks.

## 1 INTRODUCTION

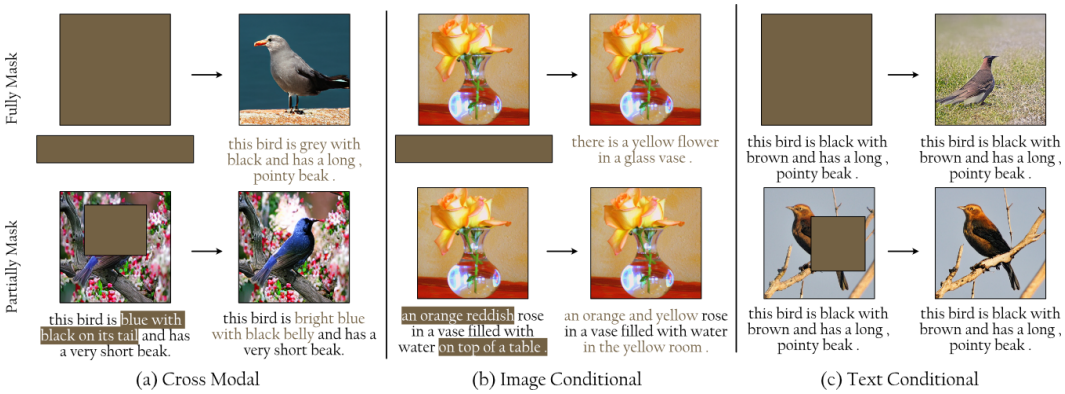



Figure 1: Examples of various tasks supported by UniD3. The dark brown portions of the image and description represent the `[MASK]`.

Diffusion models have garnered significant interest on various high quality image generation tasks, such as image generation (Ho et al., 2020), image super-resolution (Rombach et al., 2022), image inpainting (Lugmayr et al., 2022), image editing (Avrahami et al., 2022), image translation (Saharia et al., 2022a), among others. Concurrently, the Vector Quantized (VQ) models have also achieved rapid advances in image generations, especially on cross-modal tasks, examples include text-to-image (Kim et al., 2022), sketch-to-image (Esser et al., 2021b), image-to-video (Wu et al., 2021). In this paper, we are interested in a specific form of discrete diffusion model, which is designing a vision-language cross-modal framework. The discrete diffusion models have a wide range of applications in the latest emerged works, such as VQ-Diffusion (Gu et al., 2022) and Latent Diffusion Models (LDM) (Rombach et al., 2022). However, exiting works often require conditioning the generation on another input signal, either training the model for text-to-image, or taming a model for image-to-text (image caption).

Here, we take the use of discrete diffusion model into a new realm - *cross-modal* generation using a *unified vision-language model*. Specifically, we show that the internal relationships of sequences

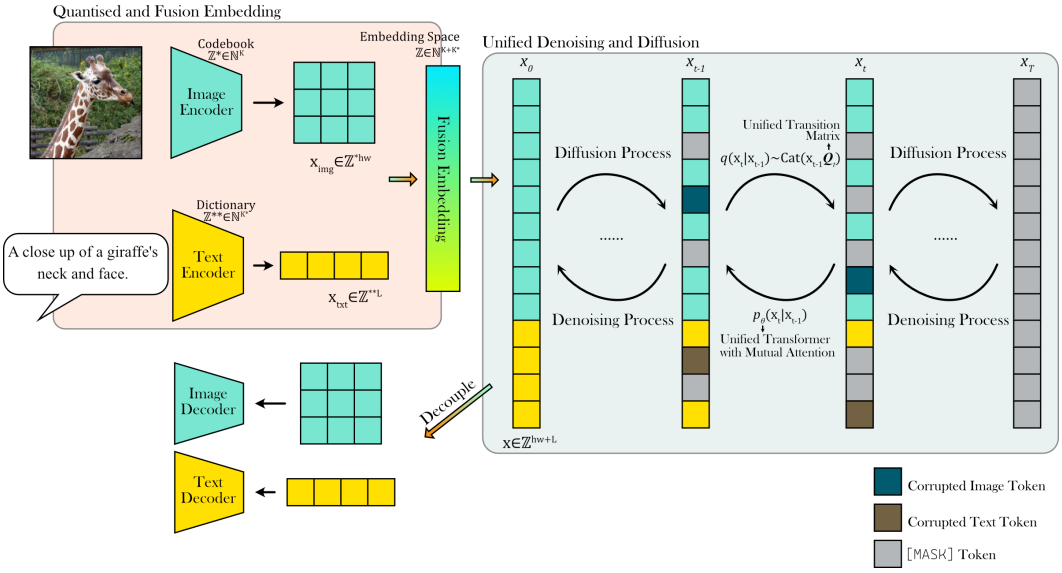



Figure 2: The pipeline of UniD3. With an offline model (red background part), the given inputs are represented by discrete token sequence in separate domain. The fusion embedding concatenate the tokens in different modal and embed them to the same space. The unified diffusion (in blue background) will construct the joint distribution of all modalities based on the fused embedding with a fixed unified Markov transition matrix.

within the vision-language pairs typically results in learning a powerful joint generative model. *UniD3*, our new **Uni**fied **D**iscrete **D**enoising **D**iffusion model, allows us to construct a joint vision-language probability distribution by mixing discrete image tokens with text tokens, leading to a capability of simultaneously generating *cross-domain* results. This is achieved by the two-stages framework, illustrated in Fig. 2, (1) An offline model to generate a compact yet expressive discrete representation for both images and texts via discrete VAE (dVAE) and Byte-Pair Encoding (BPE), respectively (the pink part in Fig. 2). (2) A novel unified discrete diffusion model to estimate the joint distribution of such latent visual and language codes (the cyan part in Fig. 2). Once trained, *UniD3* can not only inherit the ability to manipulate the provided text or image, but is also able to unify the text and image generation, *e.g.*, unconditional vision-language pairings generation, cross modal manipulation, text guided image completion, and image conditional text caption (Fig. 1 depicts the tasks that our model is capable of handling).

Based on the empirical exploration, our model can achieve comparable image quality to the text condition in the unconditional case. In terms of image caption, our model is also comparable to some states of the art.

In summary, our key contributions include the following:

- We design a specific Markov transition matrix for our unified discrete denoising diffusion model to estimate the joint distribution of language and image. We also propose mutual mechanism with fuse embedding to achieve the goal of multi-modal integration.

- The proposed UniD3 pipeline also allows multi-modal discrete representation as input. In addition, our model is able to capture the inner relations between different modalities and explore the combination within various modalities with broader freedom.

- To the best of our knowledge, the UniD3 is the first work to perform simultaneous unconditional vision-language generation and bi-directional vision-language synthesis tasks with a single model. And our methodology may potentially accept more diverse modalities than linguistic and visual.

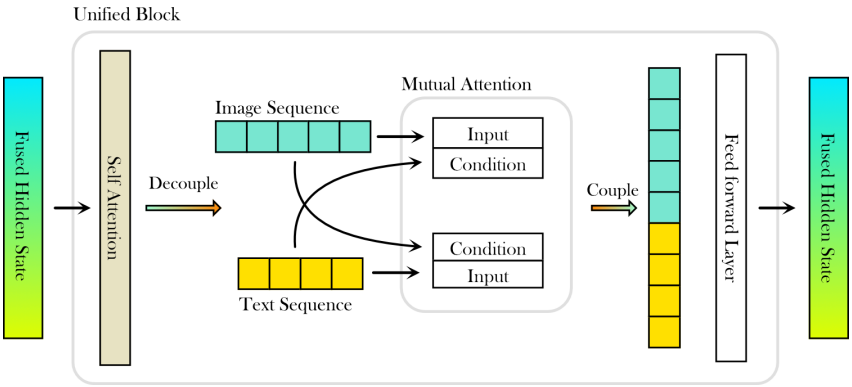



Figure 3: Illustration of transformer blocks with mutual attention. A unified transformer is composed of several blocks stacked on top of one another.

## 2 PRELIMINARIES

**Vector Quantised Model** The Vector-Quantised Variational Auto Encoder (VQ-VAE) (Van Den Oord et al., 2017) learns to embed the high-dimensional data, e.g., image or audio, into a discrete representation. In particular, given a high dimensional data $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$, the encoder $E$ first converts it to the spatial latent features $\boldsymbol{z} = \{z_{i,j}\} \in \mathbb{R}^{d\times h\times w}$, and then transfer this continuous features into discrete space by looking up the closest features in the codebook $\mathcal{Z} = \{z_k\} \in \mathbb{R}^{K\times d}$ to obtain the tokens $\boldsymbol{z}_q$:

$$\boldsymbol{z}_q = \text{Quantise}(\boldsymbol{z}) = \text{Quantise}(E(\boldsymbol{x})) := \arg\min_k ||z_{i,j} - z_k||, \tag{1}$$

where the dimensions $h, w$ of latent feature $\boldsymbol{z}$ are substantially smaller than the original dimensions $H, W$. The reconstructions can be obtained through a decoder $G$:

$$\hat{\boldsymbol{x}} = G(\boldsymbol{z}_q) = G(\text{Quantise}(E(\boldsymbol{x}))). \tag{2}$$

Recently, there have been significant advances in learning more compact representation, yet higher-quality reconstruction, such as introducing new losses (Esser et al., 2021a), applying powerful backbone (Yu et al., 2021), using multiple channels representation (Lee et al., 2022) and spatially normalization (Zheng et al., 2022). However, in this paper, we focus on applying the technology in our novel UniD3, rather than exploring a new codebook learning approach.

**Discrete Diffusion** The discrete diffusion model was originally mentioned in (Sohl-Dickstein et al., 2015), with transitions converging to a binomial distribution. Subsequent work extended this to multinomial diffusion (Hoogeboom et al., 2021; Song et al., 2021), while Austin et al. (2021) provided more options for transition matrices. Recent works integrated discrete diffusion models with VQ-VAE, allowing for high-quality image synthesis (Gu et al., 2022; Hu et al., 2022).

Here we briefly describe the multinomial diffusion with the absorbing state, as employed in VQ-Diffusion (Gu et al., 2022). Besides $K$ tokens from a discrete VAE, an additional `[MASK]` token is introduced. The forward process is defined as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \text{Cat}(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{Q}_t\boldsymbol{x}_{t-1}) = \boldsymbol{x}_t^T\boldsymbol{Q}_t\boldsymbol{x}_{t-1}, \tag{3}$$

where $\boldsymbol{x}$ is a one-hot vector identifying the token index. Here $[\boldsymbol{Q}_t]_{i,j} = q(x_t = i|x_{t-1} = j) \in \mathbb{R}^{(K+1)\times(K+1)}$ is the Markov transition matrix from $t-1$ to $t$, which can be expressed as:

$$\boldsymbol{Q}_{[t-1\to t]} = \begin{bmatrix} \alpha_t+\beta_t & \beta_t & \beta_t & \cdots & 0 \\ \beta_t & \alpha_t+\beta_t & \beta_t & \cdots & 0 \\ \beta_t & \beta_t & \alpha_t+\beta_t & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_t & \gamma_t & \gamma_t & \cdots & 1 \end{bmatrix}, \tag{4}$$

where $\alpha_t \in [0, 1]$ is the probability of retaining the token, and has a probability of $\gamma_t$ to be replaced by the `[MASK]` token, leaving a chance of $\beta_t = (1 - \alpha_t - \gamma_t)/K$ to be diffused.

The posterior of the diffusion process can be formulated as:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} = \frac{(\boldsymbol{x}_t^T \boldsymbol{Q}_t \boldsymbol{x}_{t-1})(\boldsymbol{x}_{t-1}^T \overline{\boldsymbol{Q}}_{t-1} \boldsymbol{x}_0)}{\boldsymbol{x}_t^T \overline{\boldsymbol{Q}}_t \boldsymbol{x}_0}, \tag{5}$$

given $\overline{\boldsymbol{Q}}_t = \boldsymbol{Q}_t \cdots \boldsymbol{Q}_1$, which can be calculated in closed form, and

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathrm{Cat}(\boldsymbol{x}_t; \boldsymbol{p} = \overline{\boldsymbol{Q}}_t \boldsymbol{x}_0) = \boldsymbol{x}_t^T \overline{\boldsymbol{Q}}_t \boldsymbol{x}_0. \tag{6}$$

In the reverse process, instead of explicitly predicting the posterior using a denoising neural network, the $\boldsymbol{x}_0$-parameterisation might increase the stability and permit fast inference (skipping $\Delta t$ steps per iteration). The reverse transition with reparameterisation is given as:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \propto \sum_{\tilde{\boldsymbol{x}}_0} q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0)\tilde{p_\theta}(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t), \tag{7}$$

in which the neural network predicts the logits of the target data $q(\boldsymbol{x}_0)$.

## 3 METHOD

Our goal is to adapt the discrete diffusion model to learn the joint distribution of linguistic and visual features concurrently. First, we propose a transition matrix that allows the diffusion model to capture the implicit association between text and images and the corresponding objective function. Second, we present a mutual attention transformer architecture with a fused embedding method as the denoising function, which fits our unified diffusion process objective and permits for more precise predictions. Finally, we demonstrate the possibility of extending the model to more modalities.

Our overall pipeline is illustrated in Fig. 2. Specifically, our solution begins by compressing the figures and texts into discrete token sequences using dVAE and BPE, respectively. A robust diffusion model with unified transition matrix is then constructed to fit the joint distribution across different modalities, which is further empowered by a transformer with a mutual attention mechanism.

### 3.1 UNIFIED DIFFUSION PROCESS

**Unified Transition Matrix** Discrete diffusion models with transition matrices can capture global links. The presence of a transition matrix determines the nature of the discrete diffusion model, which also provides us with more choices for token evolution. Thus we may wonder if it is feasible to design transition matrices that capture the global connections between various modalities.

The Markov transition matrix of the discrete diffusion model should satisfy the following requirements: 1. each column in $\boldsymbol{Q}_t$ should sum to one to conserve probability mass; 2. each column in the cumulative-product $\overline{\boldsymbol{Q}}_t$ should converge to either a known stationary distribution or a learnt prior when $t$ becomes large.

On the basis of these criteria, we construct a unified transition matrix $[\boldsymbol{Q}_t]_{i,j} = q(x_t = i|x_{t-1} = j)$ capable of encapsulating discrete representations among various modalities. The following matrix $\boldsymbol{Q}_t \in \mathbb{R}^{(K+K^*+1)\times(K+K^*+1)}$ illustrates a unified transition process with only text and image modalities:

$$\boldsymbol{Q}_t = \left[\begin{array}{ccccc|ccccc|c}
\alpha_t+\beta_t & \beta_t & \beta_t & \cdots & \beta_t & 0 & 0 & 0 & \cdots & 0 & 0 \\
\beta_t & \alpha_t+\beta_t & \beta_t & \cdots & \beta_t & 0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\beta_t & \beta_t & \beta_t & \cdots & \alpha_t+\beta_t & 0 & 0 & 0 & \cdots & 0 & 0 \\
\hline
0 & 0 & 0 & \cdots & 0 & \alpha_t+\beta_t^* & \beta_t^* & \beta_t^* & \cdots & \beta_t^* & 0 \\
0 & 0 & 0 & \cdots & 0 & \beta_t^* & \alpha_t+\beta_t^* & \beta_t^* & \cdots & \beta_t^* & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & \beta_t^* & \beta_t^* & \beta_t^* & \cdots & \alpha_t+\beta_t^* & 0 \\
\hline
\gamma_t & \gamma_t & \gamma_t & \cdots & \gamma_t & \gamma_t & \gamma_t & \gamma_t & \cdots & \gamma_t & 1
\end{array}\right], \tag{8}$$

where $\alpha_t \in [0, 1]$ is the probability to keep this token, $\beta_t$ and $\beta_t^*$ are the probabilities of a token to be replaced by any other accessible tokens in different modality, and $\gamma_t$ is the absorbing probability.

The dimension of the matrix $\boldsymbol{Q}_t$ is $(K + K^* + 1) \times (K + K^* + 1)$, where $K$ and $K^*$ are the number of states in different modals respectively, e.g., $K$ is the size from the codebook in discrete VAE and $K^*$ is the dictionary size of BPE.

The matrix comprises five sections:

- The final row and column form a section associated with the transition of the absorbing state. Intuitively, if the token is `[MASK]` at $t-1$ step, then the token must be `[MASK]` at time $t$. Conversely, any other modal token has the equal possibility $\gamma_t$ of being diffused to `[MASK]`.
- The remainder of the matrix is composed of four quadrants, the first and third of which are fully zeros. Specifically, these two sub-matrices prevent tokens transitioning from one modality to another. The form of the second and fourth quadrants closely resembles that for multinomial diffusion. Here, in addition to some probability of being converted into a `[MASK]` token, each token also has some chance of transiting to a different state within the same modality, or remaining unaltered.
- The dimensions of these four quadrants are not identical, which are $K^* \times K$, $K \times K$, $K \times K^*$ and $K^* \times K^*$, respectively.

It is worth noting that $\alpha_t$ and $\gamma_t$ are the same in all modalities, whereas $\beta_t$ varies based on the number of states in different modalities. Mathematically, $\beta_t = (1-\alpha_t-\gamma_t)/K$ and $\beta_t^* = (1-\alpha_t-\gamma_t)/K^*$.

The sum of each column in this transition matrix is one to preserve probability mass, and also all the mass of the stationary distribution falls on the `[MASK]` token, which satisfies the prerequisite for a discrete diffusion model transition matrix.

The computation of $\overline{\boldsymbol{Q}}_t \boldsymbol{x}_0$, needed for $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ in Eq. 6, can be efficiently obtained in closed form:

$$\begin{aligned}
&\overline{\boldsymbol{Q}}_t \boldsymbol{x}_0 = \overline{\alpha}_t \boldsymbol{x}_0 + \left[\overline{\gamma}_t - \mathbb{1}(\boldsymbol{x}_0)\overline{\beta}_t - \mathbb{1}^*(\boldsymbol{x}_0)\overline{\beta}_t^*\right] \boldsymbol{x}_{[\mathtt{M}]} + \mathbb{1}(\boldsymbol{x}_0)\overline{\beta}_t + \mathbb{1}^*(\boldsymbol{x}_0)\overline{\beta}_t^*, \\
&\text{where } \mathbb{1}(\boldsymbol{x}_0) = \begin{cases} 1 & \text{if } \operatorname{argmax} \boldsymbol{x}_0 \in [0, K), \\ 0 & \text{otherwise.} \end{cases}, \mathbb{1}^*(\boldsymbol{x}_0) = \begin{cases} 1 & \text{if } \operatorname{argmax} \boldsymbol{x}_0 \in [K, K+K^*), \\ 0 & \text{otherwise.} \end{cases}, \\
&\boldsymbol{x}_{[\mathtt{M}]} = \boldsymbol{x} \leftarrow \operatorname{argmax} \boldsymbol{x} = K + K^* \text{and } \overline{\alpha}_t, \overline{\beta}_t, \overline{\beta^*}_t, \overline{\gamma}_t \text{ are the corresponding cumulative product.}
\end{aligned} \tag{9}$$

The detailed proof is provided in Appendix D.

**Unified Objective** For conditional generation tasks such as text-to-image generation, the goal is to maximize $p(\boldsymbol{x}|\boldsymbol{y})$ by finding $\boldsymbol{x}$, where $\boldsymbol{y}$ is the given text condition. In our task, it can be approximated that the model try to maximize the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ simultaneously. In practice, we minimize the Kullback-Leibler divergence between $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ in both image and text directions, as shown in the following:

$$\mathcal{L}_0 = -\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0^{img}|\boldsymbol{x}_1, \boldsymbol{x}_0^{txt}) + \log p_\theta(\boldsymbol{x}_0^{txt}|\boldsymbol{x}_1, \boldsymbol{x}_0^{img})], \tag{10}$$

$$\mathcal{L}_{t-1} = \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[ D_{\mathrm{KL}} \left( q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \| \left[ p_\theta(\boldsymbol{x}_{t-1}^{img}|\boldsymbol{x}_t); p_\theta(\boldsymbol{x}_{t-1}^{txt}|\boldsymbol{x}_t) \right] \right) \right], \tag{11}$$

where $p_\theta(\boldsymbol{x}_{t-1})$ is the integration of the logits $p_\theta(\boldsymbol{x}_{t-1}^{img})$ and $p_\theta(\boldsymbol{x}_{t-1}^{txt})$ from separate modal, and they can obtained with $\boldsymbol{x}_0$-parameterisation as following:

$$p_\theta(\boldsymbol{x}_{t-1}^{img}|\boldsymbol{x}_t) \propto \sum_{\tilde{\boldsymbol{x}}_0^{img}} q(\boldsymbol{x}_{t-1}^{img}|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0^{img}, \boldsymbol{x}_0^{txt}) \tilde{p_\theta}(\tilde{\boldsymbol{x}}_0^{img}|\boldsymbol{x}_t)], \tag{12}$$

$$p_\theta(\boldsymbol{x}_{t-1}^{txt}|\boldsymbol{x}_t) \propto \sum_{\tilde{\boldsymbol{x}}_0^{txt}} q(\boldsymbol{x}_{t-1}^{txt}|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0^{txt}, \boldsymbol{x}_0^{img}) \tilde{p_\theta}(\tilde{\boldsymbol{x}}_0^{txt}|\boldsymbol{x}_t)]. \tag{13}$$

Due to the $x_0$-parameterisation, our model is also capable of achieving fast sampling by increase the step size. And the last term of the variational lower bound loss $\mathcal{L}_T$ is a constant and can be ignored during the training:

$$\mathcal{L}_T = D_{\mathrm{KL}}\left(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T)\right), \tag{14}$$

as the prior $p(\boldsymbol{x}_T)$ is fixed with the unified transition matrix:

$$p(\boldsymbol{x}_T) = \left[\overline{\beta}_T, \overline{\beta}_T, \ldots, \overline{\beta}^*_T, \overline{\beta}^*_T, \ldots, \overline{\gamma}_T\right]. \tag{15}$$

The full expression of the loss function can be found in Appendix B.

### 3.2 Denoising Function for Multimodal

**Mutual Attention** As indicated by Eqs. 12 & 13, the neural network is responsible for prediction of the distribution $\tilde{p_\theta}(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t)$. However, the input to the neural network covers all modalities, and a simple self-attention mechanism can scarcely highlight the inter-modal linkages. In addition, the distributions predicted by the neural network need to be decoupled according to the various modalities throughout the reparameterisation. In other words, the network predictions should be expressed in terms of different modalities, e.g., $\tilde{p_\theta}(\tilde{\boldsymbol{x}}_0^{img}|\boldsymbol{x}_t)$ and $\tilde{p_\theta}(\tilde{\boldsymbol{x}}_0^{txt}|\boldsymbol{x}_t)$, with interconnections.

Therefore, we propose the mutual attention mechanism and construct the unified transformer as shown in Fig. 3. The unified transformer contains several transformer blocks, each of which consists of one self-attention and two parallel mutual attention operations, and one feed-forward module. Each block receives a sequence of mixed-modal tokens as input that traverses a layer of self-attention to capture the inherent connection within the modalities. Afterwards, the hidden tokens are decoupled according to the location of the different modalities and fed to the corresponding mutual attention layers. Mutual attention is a modified version of the cross-attention layer, with the conditional inputs to cross-attention being maintained constant while the inputs to mutual attention are derived from hidden features. Next, the outputs from the various mutual attention are concatenated into one mixed-modal token sequence for transmission to the next block. At the end of the unified transformer, we layer-normalise the sequence of tokens from the blocks, which is then decoupled to different predictive distributions by fully-connected layers.

**Fused Embedding** In addition, discrete tokens of various modalities should be embedded in the same space. Considering that two modalities have $K$ and $K^*$ states respectively, we firstly create an embedding layer with size $K + K^* + 1$, with the additional state for the `[MASK]` token. We then create a learnable spatial position encoding for the image modality and a learnable sequence position encoding for the text modality. The final fused embedding is obtained by adding the embedded vectors to the positional encoding of their associated modalities.

### 3.3 Extension to More Modalities

The above description emphasizes textual and visual modalities. However, our proposed methodology is also applicable to other discretisable modalities. Even within the visual domain, we may further consider distinct modalities such as bounding boxes, segmentation masks and edge maps, in addition to RGB images. Specifically, our transition matrix in Eq. 8 is extensible, we can have additional modalities by simply adding new modal quadrants and leaving the mask transition in the final row and column intact. Should the number of states in a modality become excessive, it will degrade to an absorbing diffusion. It is straightforward that the extended unified transition matrix also meets the criteria of the discrete diffusion model, with the stationary distribution assigning all probability mass to the `[MASK]` token.

The unified objective can also include more modalities. When there are more modalities, $p_\theta(\boldsymbol{x}_{t-1})$ consists of more distributions, e.g., $p_\theta(\boldsymbol{x}^0_{t-1}), p_\theta(\boldsymbol{x}^1_{t-1}), \cdots, p_\theta(\boldsymbol{x}^n_{t-1})$. The architecture of the neural network used for prediction $\tilde{x}_0$ can also be altered by adding the mutual attention to corresponding modal, which is conditioned on the remaining modal sequences. Due to the fact that the complexity of the transformer is quadratic to the length of the sequence, excessive modalities may incur enormous memory cost.

The above demonstrates the ability of the proposed model to handle complex modalities; however we leave the exploration and optimisation of more modalities to future work.

Figure 4: Generated vision-language Pairs from CUB-200 and MSCOCO. Both the image and caption are *generated simultaneously.* The quality of the created photos and text is comprehensible, and there is a correlation between the descriptions and the visuals.

## 4 Experiments

The description of the datasets and model experimental details can be found in Appendix A.

### 4.1 Quantitative results

In the unconditional scenario, our model is able to sample from all modalities while maintaining the relationship between them. For conditional generation, we conducted experiments on text-based image generation and image-based caption generation, respectively.

**Metrics** We evaluate the multimodal capacity of our model in three distinct areas: Image sampling quality, text sampling quality and the similarity between sampled image and text. Fréchet inception distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016) are used to evaluate the realism and variety of the produced images, while perplexity (PPL) with GPT-2 is used to evaluate the quality of unconditional text samples. The BLEU-4, METEOR and SPICE are used to assess the quality of a picture caption. And CLIP (Radford et al., 2021) is used to identify the similarity between visuals and text.

**Image Quality** We quantitatively compare our unconditional and text-conditional results with several state-of-the-art text-to-image and bidirectional conditional generation methods, including GAN-based methods, diffusion-based methods and autoregressive methods. The comparison results are given in Table 1, in which *Pair* means the unconditional generated vision-language pairings and *T2I* represents the text-to-image generation.

The first section of the table solely covers text-to-image models, with the exception of LAFITE (Zhou et al., 2022), a language-free model. All the results in this part are training on the MSCOCO datasets from scratch except LDM (Rombach et al., 2022) which is trained on LAION-400M and conducts zero-shot text-to-image generation based on MSCOCO evaluation captions. The results of LDM is with a $f8$-KL based image encoder and fast sampling strategy, excluding the classifier guidance during inference. In the second part, these models are claimed to be endowed with cross modal capabilities. However, such methods are not able to achieve simultaneous vision-language pair generation. For a fair comparison, we exhibit the quantitative generative results based on various modalities individually. We provide more experimental results for text conditional generation in Appendix E.2

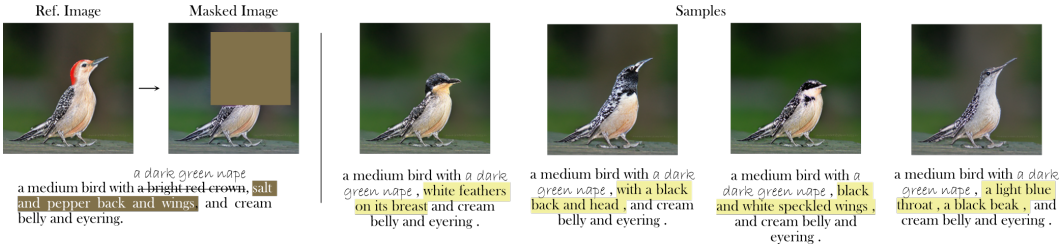



Figure 5: Presentation of the results of cross modal vision and language manipulation and infilling. The dark brown portions of the image and description represent the [MASK], while the strike-through represent the caption manipulation. Image and text compliment each other simultaneously.

| | CUB | | MSCOCO | |
|---|---|---|---|---|
| | FID (↓) | IS (↑) | FID (↓) | IS (↑) |
| DM-GAN (Zhu et al., 2019) | 16.09 | 4.75 | 32.64 | 27.3 |
| DF-GAN (Tao et al., 2022) | 14.81 | 5.1 | 21.42 | - |
| VQ-Diffusion (Gu et al., 2022) | 11.94 | - | 19.75 | - |
| LDM w/o G (Rombach et al., 2022) | - | - | 23.31 | 20.23 |
| LAFITE(LF) (Zhou et al., 2022) | 27.53 | 4.32 | 18.04 | 27.2 |
| Unifying (Huang et al., 2021) | - | - | 29.9 | - |
| L-Verse (Kim et al., 2022) | - | - | 45.8 | - |
| OFA (Wang et al., 2022) | - | - | 10.5 | 31.1 |
| Ours(Pair) | 17.38 | 6.11 | 28.63 | 24.44 |
| Ours(T2I) | 16.19 | 6.02 | 25.11 | 21.03 |

Table 1: Evaluation of image synthesis on the CUB-200-2011 and MS-COCO dataset. The first section is the text-conditioned results while the model in second part includes multimodal solution.

| | B-4 (↑) | M (↑) | S (↑) |
|---|---|---|---|
| BUTD (Anderson et al., 2018) | 36.3 | 27.7 | 21.4 |
| ORT (Herdade et al., 2019) | 38.6 | 28.7 | 22.6 |
| AoA (Huang et al., 2019) | 38.9 | 29.3 | - |
| X-LAN (Pan et al., 2020) | 39.7 | 29.5 | 23.4 |
| SimVLM (Wang et al., 2021) | 40.6 | 33.7 | 25.4 |
| OSCAR (Li et al., 2020) | 40.5 | 29.7 | 22.8 |
| Unifying (Huang et al., 2021) | 37.3 | 28.2 | 21.9 |
| L-verse (Kim et al., 2022) | 39.9 | 31.4 | 23.3 |
| OFA (Wang et al., 2022) | 41.0 | 30.9 | 24.2 |
| Ours(I2T) | 39.6 | 29.3 | 23.4 |

Table 2: Evaluation of image caption on the MSCOCO *Karpathy* split. The first section is the text-conditioned results while the model in second part includes multimodal solution.

**Text Quality** In the case of image captions, we compared several other text description methods, the details are shown in Table 2, where *I2T* is the image-based text-generation task. Similarly, the first part of the table has pure image caption solutions and the bi-directional methods are demonstrated in the second part. Some results for image caption tasks are provided in Appendix E.1

It is worth noting that the caption generation in our *Pair* scenario is unique. Common text generation metrics are difficult to implement due to a lack of the references. Here we used Perplexity to evaluate the text quality. The perplexities of generated CUB-200 and COCO captions are 123.32 and 188.74 whereas the corresponding values for the evaluation set are 108.66 and 175.36.

**Similarity** In the multimodal generation phase, all the other methods for comparison need the input of at least one specified modality; instead our method generates both visual and verbal features. The generated vision-language pairings are shown in Fig. 4 while more pair samples are provided in Appendix E.4. In this experiment, we use CLIP scores to evaluate the similarity between our generated visual and verbal modalities, while generated images with given texts were used for other methods. The comparison results are given in Table 3, where a higher correlation between the text/image samples of our model can be found.

### 4.2 Cross Modal Manipulation

To demonstrate the modification potential of our vision-language federation, we present the outcomes of picture alteration with description modification in Fig. 5 and more samples can be found in Appendix E.3. In this experiment, we obscured a portion of the image and modified the text description as well. We anticipate that the model will complete the image based on the amended text, in conjunction with the description based on the visible image region.

Depending on the unmasked picture, our model may supplement the caption that is derived from a portion of the image. Moreover, based on the amended description, the model may enhance the appearance of the masked component. These findings indicate that our model is capable of modifying text and pictures in both ways, in addition to jointly generating text and images.

### 4.3 Ablation Study

In the ablation test, we utilized a rapid sampling scheme, with an interval between each step of 10. First, we evaluated the outcomes without mutual attention, where we replaced all mutual attention in the model structure with causal self-attention. In addition, we evaluated the performance without the unified transition matrix, we employed a matrix similar to Equation 4, with the out-of-bound outputs substituted by the minimal ordinal number of the corresponding modal.

| (↑) | CUB | COCO |
|---|---|---|
| Eval set | 0.286 | 0.306 |
| VQ-Diff (Gu et al., 2022) | - | 0.257 |
| iVQ-Diff (Tang et al., 2022) | - | 0.304 |
| Unifying (Huang et al., 2021) | - | 0.309 |
| OFA (Wang et al., 2022) | - | 0.344 |
| Ours(Pair) | 0.280 | 0.301 |
| Ours(T2I) | 0.293 | 0.306 |
| Ours(I2T) | 0.302 | 0.312 |

Table 3: The CLIP similarity between the generated captions and images.

| (↓) | FID |
|---|---|
| Ours | 17.38 |
| *w/* Fast Sampling $\Delta$ t = 10 | 20.96 |
| *w/o* Mutual Attention | 23.19 |
| *w/o* Unified Transition Matrix | 32.63 |
| Ours (T2I) | 16.19 |
| *w/* Fast Sampling $\Delta$ t = 10 | 18.33 |
| *w/o* Mutual Attention | 21.65 |
| *w/o* Unified Transition Matrix | 28.93 |

Table 4: Ablations studies. The experiments are conducted on the CUB-200.

## 5 Related Work

**Denoising Diffusion Model** Sohl-Dickstein et al. (2015) initially presented the diffusion model, which consists of a base Gaussian diffusion and a basic binomial diffusion. Ho et al. (2020) recently reignited interest in the extraordinary generating capacity of diffusion models. This is extended in subsequent research to enhance quality with the classifier guidance (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021) and speed of sampling (Song et al., 2021; Salimans & Ho, 2021). In addition to continuous diffusion models that have received the significant attention, there is also substantial research based on discrete diffusion models (Hoogeboom et al., 2021; Austin et al., 2021) for text generation or image segmentation. With the help of VQ-VAE, Esser et al. (2021a); Hu et al. (2022); Gu et al. (2022); Tang et al. (2022) discrete diffusion models have become capable of generating high quality images. Diffusion models can also capture the connection between the visual and textual embedding and model a proper distribution (Nichol et al., 2021; Ramesh et al., 2022; Gu et al., 2022; Saharia et al., 2022b).

**multimodal Generation** Text-to-image generation is one of the key aspects of multimodal generation. On simpler datasets, traditional GAN-based models (Xu et al., 2018; Zhu et al., 2019) may create high-quality pictures based on text hints, but face difficulties on more complex datasets. As a result of the emergence of transformers employing the attention mechanism (Vaswani et al., 2017), conventional autoregressive models (Van den Oord et al., 2016) have gained more powerful generating capacities as ART (Auto-regressive Transformers) (Ramesh et al., 2021; Esser et al., 2021b; Yu et al., 2022). Other types of multimodal generation involve the addition of prospective modalities as criteria for improved generation outcomes (Gafni et al., 2022). There are also some works that focus on addressing bidirectional generation between different modalities (Kim et al., 2022; Huang et al., 2021; Wang et al., 2022); however, these solutions are task-specific, necessitating different input formats and models for tasks with distinct modalities.

## 6 Conclusion

This work presents UniD3, a novel unified framework for simultaneous multimodal generation. We introduced a unified transition matrix for the diffusion process which allows the states of fused tokens to be corrupted in a stable way. To capture the connection from the various modalities, we suggested a fusion embedding and mutual attention for noiseless state recovery. Our approach is also general and capable of the conditional generation problem for the given modality. The results on sophisticated datasets are intriguing for further exploring, as well as the performance of more modalities is also one of the future directions.

## 7 ETHICS

In this work, we propose UniD3, a new method for multimodal multi-task generation based on the discrete denoising diffusion model. All datasets evaluated in our experiments are open source, publicly accessible, and used with permission.

Similar to other generation algorithms, our method has both beneficial and negative societal effects, depending on its application and usage.

- Positively, UniD3 may explore the limits of deep learning generation through a unique pattern of graphic pair generation, while letting the typical user's imagination run wild and decreasing the entrance barrier for visual content production.
- On the other hand, UniD3 may be used to make modified photographs of great quality that are difficult to discern from the originals. Ways to fool humans and propagate falsehoods that are malicious. UniD3 can be abused by nefarious users, which may have serious repercussions.
- In our code distribution. We shall outline the permitted uses of our system and give the corresponding licences.

However, we observe that the present discriminators used to distinguish generated images are ineffective at identifying images generated by diffusion models. There is still the necessity for more exploration to discern between authentic and fake images.

## 8 REPRODUCIBILITY

In order to achieve reproducibility, we have made the following efforts:

1. We have submitted an original repository along with the test code as supplementary materials. The file also contains an anonymous link to a downloadable model.
2. Our tidy codes will be released upon publication.
3. Details on dataset and model architectures are provided in Appendix A, as well as more experimental settings are provided in Appendix C.
4. Some critical proofs are included in Appendix B & D.
5. Additional experimental results are provided in Appendix E.

## REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

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

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021a.

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

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and Ponnuthurai N Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11502–11511, 2022.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4634–4643, 2019.

Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1138–1147, 2021.

Taehoon Kim, Gwangmo Song, Sihaeng Lee, Sangyun Kim, Yewon Seo, Soonyoung Lee, Seung Hwan Kim, Honglak Lee, and Kyunghoon Bae. L-verse: Bidirectional generation between image and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16526–16536, 2022.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10971–10980, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022b.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021.

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

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16515–16525, 2022.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2021.

Wansen Wu, Tao Chang, and Xinmeng Li. Visual-and-language navigation: A survey and taxonomy. *arXiv preprint arXiv:2108.11544*, 2021.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *International Conference on Learning Representations*, 2021.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35, 2022.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17907–17917, 2022.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5810, 2019.

# A DATASETS AND MODEL DETAILS

We demonstrate the feasibility of the proposed method on two commonly used datasets: CUB-200 (Wah et al., 2011) and MSCOCO (Lin et al., 2014). The CUB-200 dataset consists of 8,855 training images and 2,933 test images representing 200 species of birds. Each of these photos is accompanied by 10 textual descriptions. In MSCOCO, there are 591,753 images utilized for training and 25,014 for testing, with each image corresponding to 5 textual descriptions.

We use VQ-GAN (Esser et al., 2021b) with gumbel softmax to compress the images into discrete token sequences. We directly use the pre-trained and public model which is trained on OpenImage (Krasin et al., 2017). The compression ratio of the model is $8 \times 8 \times 3 = 192$, and each image is compressed into a $32 \times 32$ token sequence. The codebook size is 2,887 after removing useless codes (Gu et al., 2022).

Furthermore, as the transition matrix of the diffusion model is sensitive to the number of dictionary size and the length of the sequence influences the performance of the denoising neural network, we further investigated the effect of various forms of text word lists. We select the dVAEs with different downsampling factors and dictionary size from (Esser et al., 2021b). We used alternative models of VQ-GAN release, including different downsampling factor with different codebook sizes, from $f8$ with 8,192 terms $\mathcal{Z} \in \mathbb{R}^{8192\times256}$ to $f16$ with 16,384 entries $\mathcal{Z} \in \mathbb{R}^{16384\times256}$.

| (↓) | FID | Sample Steps |
|---|---|---|
| Ours | 17.38 | 500 |
| *w/* $f8$-8192 Gumbel Image Encoder | 25.68 | 100 |
| *w/* $f16$-974 Image Encoder | 21.02 | 100 |
| *w/* $f16$-16384 Image Encoder | 24.61 | 100 |

Table 5: Ablations studies on Image Encoder. The experiments are conducted on the CUB-200.

For the text portion, we use a BPE tokenizer comparable to (Ramesh et al., 2021; Radford et al., 2021) and set the dictionary size to 8,192, compressing each text description into a sequence of length 128. The text and image encoders are fixed in the training phase.

For the diffusion process, we set the number of diffusion steps to 500. And the noise planning is linear, where $\overline{\alpha_t}$ goes from 1 to 0 and $\overline{\gamma_t}$ goes from 0 to 1. The denoising network architecture is as described in Sec.3.2, in which the transformer comprises 20 transformer blocks with 16 heads attention and 1024 feature dimension. The model contains 600M parameters. For the ablation model, we use 18 transformer blocks with 16 heads, and the dimension is 256. The model contains 119M parameters. The optimiser for the model is AdamW (Loshchilov & Hutter, 2018), and the learning rate is $9e^{-4}$ without warmup. We trained all models with a batch size of 16 across 8 Tesla A100.

# B LOSS FUNCTION

Similar to the continuous domain, the complete expression of the loss function in our model is:

$$\begin{aligned}\mathcal{L}_{vb} =& \mathbb{E}_{q(\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T)\right)\right] \\ &+ \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|\left[p_\theta(\boldsymbol{x}_{t-1}^{img}|\boldsymbol{x}_t); p_\theta(\boldsymbol{x}_{t-1}^{txt}|\boldsymbol{x}_t)\right]\right)\right] \\ &- \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0^{img}|\boldsymbol{x}_1,\boldsymbol{x}_0^{txt}) + \log p_\theta(\boldsymbol{x}_0^{txt}|\boldsymbol{x}_1,\boldsymbol{x}_0^{img})\right].\end{aligned} \tag{16}$$

# C TRUNCATION SAMPLING

As stated in (Gu et al., 2022), the truncation sampling is crucial for the discrete diffusion-based approach, which prevents the network from sampling tokens with low probability. In our experiments, the truncation rate to 0.88 for *Pair* and *T2I* generation and 0.75 for *I2T* task, mathematically, we

only keep top $88\%$ or $75\%$ tokens of $p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t)$ during inference for image and text tasks, respectively. A truncation rate that is too low will cause loss of image detail, but a rate that is too high will prohibit the image from creating a distinct geometry.

## D PROOF OF THE CLOSED-FORM SOLUTION FOR $\overline{\mathbf{Q}}_t$

The solution for $\overline{\boldsymbol{Q}}_t$ inherits the closed-form attribute from (Gu et al., 2022). Mathematically, given the initial state $\boldsymbol{x}_0$ at $t=0$, the probabilities for the next time-step $t=1$ can be obtained:

$$\overline{\boldsymbol{Q}}\boldsymbol{x}_0 = \begin{cases} \overline{\alpha}_1 + \overline{\beta}_1, & \boldsymbol{x} = \boldsymbol{x}_0 \quad \textbf{s.t.}\ \operatorname{argmax} \boldsymbol{x}_0 \in [0, K), \\ \overline{\alpha}_1 + \overline{\beta}_1^*, & \boldsymbol{x} = \boldsymbol{x}_0 \quad \textbf{s.t.}\ \operatorname{argmax} \boldsymbol{x}_0 \in [K, K+K^*), \\ \overline{\beta}_1, & \boldsymbol{x} \neq \boldsymbol{x}_0 \quad \textbf{s.t.}\ \operatorname{argmax} \boldsymbol{x}_0 \in [0, K), \\ \overline{\beta}_1^*, & \boldsymbol{x} \neq \boldsymbol{x}_0 \quad \textbf{s.t.}\ \operatorname{argmax} \boldsymbol{x}_0 \in [K, K+K^*), \\ \overline{\gamma}, & \operatorname{argmax} \boldsymbol{x} = K + K^*. \end{cases} \tag{17}$$

Suppose the closed-form expression of $\overline{\mathbf{Q}}_\tau$ holds at time-step $t=\tau$, then for the next step $t=\tau+1$:

$$\overline{\boldsymbol{Q}}_{\tau+1}\boldsymbol{x}_0 = \boldsymbol{Q}_{\tau+1}\overline{\boldsymbol{Q}}_\tau \boldsymbol{x}_0. \tag{18}$$

The outputs under different conditions can be discussed:

1. When $\boldsymbol{x} = \boldsymbol{x}_0$ and $\operatorname{argmax} \boldsymbol{x}_0 \in [0, K)$:

$$\begin{aligned} \boldsymbol{Q}_{\tau+1}\boldsymbol{x}_0 &= \overline{\beta}_\tau \beta_{\tau+1}(K-1) + (\alpha_{\tau+1} + \beta_{\tau+1})\left(\overline{\alpha}_\tau + \overline{\beta}_\tau\right) \\ &= \overline{\beta}_\tau \left(K\beta_{\tau+1} + \alpha_{t+1}\right) + \overline{\alpha}_t \left(\alpha_{t+1} + \beta_{t+1}\right) \\ &= \frac{1}{K}\left[\overline{\beta}_\tau(1-\gamma_{\tau+1}) + \overline{\alpha}_\tau\beta_{\tau+1} - \overline{\beta}_{\tau+1}) * K\right] + \overline{\alpha}_{\tau+1} + \overline{\beta}_{\tau+1} \\ &= \frac{1}{K}\left[(1-\overline{\alpha}_\tau - \overline{\gamma}_\tau)(1-\gamma_{\tau+1}) + K\overline{\alpha}_\tau\beta_{\tau+1} - (1-\overline{\alpha}_{\tau+1} - \overline{\gamma}_{\tau+1})\right] + \overline{\alpha}_{\tau+1} + \overline{\beta}_{\tau+1} \\ &= \frac{1}{K}\left[(1-\overline{\gamma}_{\tau+1}) - \overline{\alpha}_\tau(1-\gamma_{\tau+1} - K\beta_{\tau+1}) - (1-\overline{\gamma}_{\tau+1}) + \overline{\alpha}_{\tau+1}\right] + \overline{\alpha}_{\tau+1} + \overline{\beta}_{\tau+1} \\ &= \overline{\alpha}_{\tau+1} + \overline{\beta}_{\tau+1}; \end{aligned} \tag{19}$$

2. When $\boldsymbol{x} = \boldsymbol{x}_0$ and $\operatorname{argmax} \boldsymbol{x}_0 \in [K, K+K^*)$:

$$\begin{aligned} \boldsymbol{Q}_{\tau+1}\boldsymbol{x}_0 &= \overline{\beta^*}_\tau \beta^*_{\tau+1}(K^*-1) + \left(\alpha_{\tau+1} + \beta^*_{\tau+1}\right)\left(\overline{\alpha}_\tau + \overline{\beta^*}_\tau\right) \\ &= \overline{\beta^*}_\tau \left(K^*\beta^*_{\tau+1} + \alpha_{t+1}\right) + \overline{\alpha}_t \left(\alpha_{t+1} + \beta^*_{t+1}\right) \\ &= \frac{1}{K^*}\left[\overline{\beta^*}_\tau(1-\gamma_{\tau+1}) + \overline{\alpha}_\tau\beta^*_{\tau+1} - \overline{\beta^*}_{\tau+1}) * K^*\right] + \overline{\alpha}_{\tau+1} + \overline{\beta^*}_{\tau+1} \\ &= \frac{1}{K^*}\left[(1-\overline{\alpha}_\tau - \overline{\gamma}_\tau)(1-\gamma_{\tau+1}) + K^*\overline{\alpha}_\tau\beta^*_{\tau+1} - (1-\overline{\alpha}_{\tau+1} - \overline{\gamma}_{\tau+1})\right] + \overline{\alpha}_{\tau+1} + \overline{\beta^*}_{\tau+1} \\ &= \frac{1}{K^*}\left[(1-\overline{\gamma}_{\tau+1}) - \overline{\alpha}_\tau(1-\gamma_{\tau+1} - K^*\beta^*_{\tau+1}) - (1-\overline{\gamma}_{\tau+1}) + \overline{\alpha}_{\tau+1}\right] + \overline{\alpha}_{\tau+1} + \overline{\beta^*}_{\tau+1} \\ &= \overline{\alpha}_{\tau+1} + \overline{\beta^*}_{\tau+1}; \end{aligned} \tag{20}$$

3. When $\boldsymbol{x} \neq \boldsymbol{x}_0$ and $\operatorname{argmax} \boldsymbol{x}_0 \in [0, K)$:

$$\begin{aligned}
\boldsymbol{Q}_{\tau+1}\boldsymbol{x}_0 &= \overline{\beta}_\tau(\alpha_{\tau+1} + \beta_{\tau+1}) + \overline{\beta}_\tau \beta_{\tau+1}(K-1) + \overline{\alpha}_\tau \beta_{\tau+1} \\
&= \overline{\beta}_\tau(\alpha_{\tau+1} + \beta_{\tau+1}) * K + \overline{\alpha}_\tau \beta_{\tau+1} \\
&= \frac{1 - \overline{\alpha}_\tau - \overline{\gamma}_\tau}{K} * (1 - \gamma_{\tau+1}) + \overline{\alpha}_\tau \beta_{\tau+1} \\
&= \frac{1}{K}(1 - \overline{\gamma}_{\tau+1}) + \overline{\alpha}_\tau(\beta_{\tau+1} - \frac{1 - \gamma_{\tau+1}}{K}) \\
&= \overline{\beta}_{\tau+1} + \frac{\overline{\alpha}_{\tau+1}}{K} + \frac{\overline{\alpha}_\tau \left[(1 - \alpha_{\tau+1}) - 1\right]}{K} \\
&= \overline{\beta}_{\tau+1};
\end{aligned} \tag{21}$$

4. When $\boldsymbol{x} \neq \boldsymbol{x}_0$ and $\operatorname{argmax} \boldsymbol{x}_0 \in [K, K + K^*)$:

$$\begin{aligned}
\boldsymbol{Q}_{\tau+1}\boldsymbol{x}_0 &= \overline{\beta^*}_\tau(\alpha_{\tau+1} + \beta^*_{\tau+1}) + \overline{\beta^*}_\tau \beta^*_{\tau+1}(K^* - 1) + \overline{\alpha}_\tau \beta^*_{\tau+1} \\
&= \overline{\beta^*}_\tau(\alpha_{\tau+1} + \beta^*_{\tau+1}) * K^* + \overline{\alpha}_\tau \beta^*_{\tau+1} \\
&= \frac{1 - \overline{\alpha}_\tau - \overline{\gamma}_\tau}{K^*} * (1 - \gamma_{\tau+1}) + \overline{\alpha}_\tau \beta^*_{\tau+1} \\
&= \frac{1}{K^*}(1 - \overline{\gamma}_{\tau+1}) + \overline{\alpha}_\tau(\beta^*_{\tau+1} - \frac{1 - \gamma_{\tau+1}}{K^*}) \\
&= \overline{\beta^*}_{\tau+1} + \frac{\overline{\alpha}_{\tau+1}}{K^*} + \frac{\overline{\alpha}_\tau \left[(1 - \alpha_{\tau+1}) - 1\right]}{K^*} \\
&= \overline{\beta^*}_{\tau+1};
\end{aligned} \tag{22}$$

5. When $\operatorname{argmax} \boldsymbol{x} = K + K^*$:

$$\begin{aligned}
\boldsymbol{Q}_{\tau+1}\boldsymbol{x}_0 &= \overline{\gamma}_\tau + (1 - \overline{\gamma}_\tau)\gamma_{\tau+1} \\
&= 1 - (1 - \overline{\gamma}_{\tau+1}) \\
&= \overline{\gamma}_{\tau+1}.
\end{aligned} \tag{23}$$

Thus, we can obtain a closed-form solution for $\overline{\boldsymbol{Q}}_t \boldsymbol{x}_0$:

$$\overline{\boldsymbol{Q}}_t \boldsymbol{x}_0 = \begin{cases} (\overline{\alpha}_\tau + \overline{\beta}_\tau)\boldsymbol{x}_0 + (\overline{\gamma}_\tau - \overline{\beta}_\tau)\boldsymbol{x}_{[\mathtt{M}]} + \overline{\beta}_\tau, & \mathbf{s.t.}\ \operatorname{argmax} \boldsymbol{x}_0 \in [0, K), \\ (\overline{\alpha}_\tau + \overline{\beta^*}_\tau)\boldsymbol{x}_0 + (\overline{\gamma}_\tau - \overline{\beta^*}_\tau)\boldsymbol{x}_{[\mathtt{M}]} + \overline{\beta^*}_\tau, & \mathbf{s.t.}\ \operatorname{argmax} \boldsymbol{x}_0 \in [K, K + K^*). \end{cases}, \tag{24}$$

where $\boldsymbol{x}_{[\mathtt{M}]} = \boldsymbol{x} \leftarrow \operatorname{argmax} \boldsymbol{x} = K + K^*$.

Given indicator functions $\mathbb{1}$ and $\mathbb{1}^*$ for different modals:

$$\begin{aligned}
\mathbb{1}(\boldsymbol{x}_0) &= \begin{cases} 1 & \text{if } \operatorname{argmax} \boldsymbol{x}_0 \in [0, K), \\ 0 & \text{otherwise.} \end{cases} \\
\text{and } \mathbb{1}^*(\boldsymbol{x}_0) &= \begin{cases} 1 & \text{if } \operatorname{argmax} \boldsymbol{x}_0 \in [K, K + K^*), \\ 0 & \text{otherwise.} \end{cases} \\
\mathbf{s.t.} \mathbb{1}(\boldsymbol{x}_0) &\cap \mathbb{1}^*(\boldsymbol{x}_0) = \varnothing,
\end{aligned} \tag{25}$$

Eq. 24 can be expressed in:

$$\overline{\boldsymbol{Q}}_t \boldsymbol{x}_0 = \overline{\alpha}_t \boldsymbol{x}_0 + \left[\overline{\gamma}_t - \mathbb{1}(\boldsymbol{x}_0)\overline{\beta}_t - \mathbb{1}^*(\boldsymbol{x}_0)\overline{\beta}^*_t\right] \boldsymbol{x}_{[\mathtt{M}]} + \mathbb{1}(\boldsymbol{x}_0)\overline{\beta}_t + \mathbb{1}^*(\boldsymbol{x}_0)\overline{\beta}^*_t. \tag{26}$$

■

# E ADDITIONAL RESULTS

## E.1 IMAGE CAPTION

We provide some image caption results based on CUB-200 and MSCOCO dataset in Figs 6 & 7, respectively.

## E.2 TEXT CONDITIONAL GENERATION

We provide more results of pure text to image synthesis in Fig. 8. The resolution of each generated image is $256 \times 256$.

## E.3 CROSS MODAL MODIFICATION

We present samples for modification across vision and language modals under CUB-200 dataset in Fig. 9. The ref. image shows the reference image after the dVAE reconstruction.

## E.4 VISION-LANGUAGE PAIR GENERATION

A few additional examples for vision-language pair generation are shown in Fig 10. The resolution of each image is $256 \times 256$.

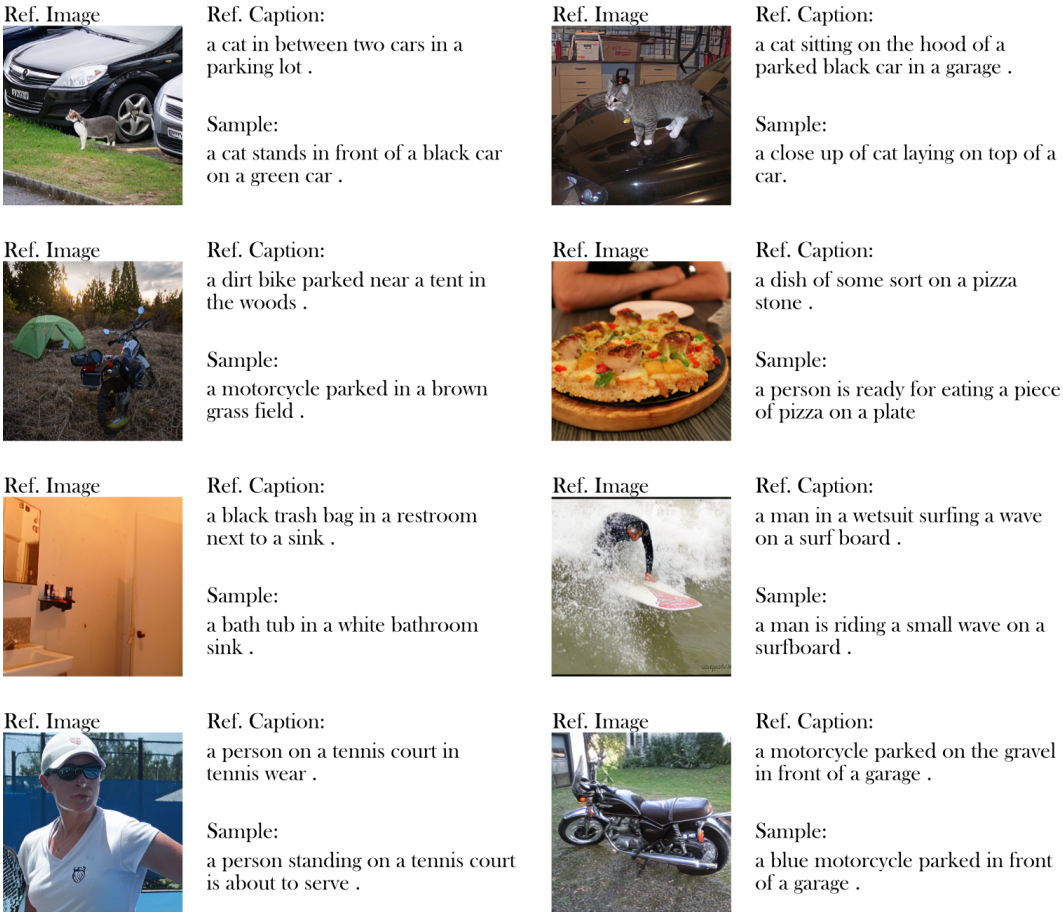



Figure 6: Samples of the image caption task based on MSCOCO dataset.

Ref. Image:

Ref. Captions:

this bird has an all black body with a large orange beak and a white eye.

the bird has a white eyering and webbed black feet.

a bird with a snubbed, rounded orange bill, stark white eyes, webbed feet, and feathered superciliary.

compared to the rest of the body, the head is relatively small, the bird has webbed feet and an orange beak.

this peculiar bird is large and mostly grey, with purple webbed feet, a bright orange, stubby beak, and a black plume on its head.

this bird is black with red and has a very short beak.

a medium sized bird with grey feathers, a black head with a white stripe leading down from the eye, a white eye, a mohawk-shaped feather on the head, and an orange beak.

the crown of the bird has a distinctive feature of black feathers protruding from the head.

this is a dark gray bird with black webbed feet, red beak and a crown that stands up

this bird with webbed feet has an orange bill, a white malar stripe, and a grayish-brown breast.

Samples:

the bird has black feathers on its cheek neck and webbed feet .

this is a dark grey bird with some dark grey wings and a large light colored beak

this big grey bird has a black body and an orange bill .

this is a black and white bird with a long , orange bill .

this ation bird has a white body , short , black beak , grey and black weand orange feet .

this bird has an orange beak with a black body gray belly and black wings

this bird has a grey body and black wings with a long orange bill .

this bird has a dark gray belly , an off orange beak and a bright orange on its eye .

Ref. Image:

Ref. Captions:

the small bird is light brown with a short beak.

this bird has a white throat, a short bill, and a gray crown.

this small bird with flat bill and a multicolored body , and black eyes.

this small bird is grey and has a small black beak

this is a gray bird with a white throat and a mohawk at the crown.

this particular bird has a gray belly and breast with black secondaries.

this thin bird has a little gray mohawk and a short tail.

this bird is grey with white and has a long, pointy beak.

this particular bird has a belly that is gray with black spots.

this bird has a brown crown with brown coverts and black beak.

Samples:

this is a small brown bird with dark brown wings grey head , and grey breast .

a small brown head and grey bird with a white belly .

small bird with brown feathers and grey back and wings

this bird has wings that are grey and white with a grey belly

this is small gray bird with an all white belly and medium colored wings with a black head .

a small bird with a small bill and a light dark gray belly and it has brown feet .

this small bird is grey and has white wings and tail , and dark gray feathers .

a small grey bird with grey darker on its wings .

Figure 7: More samples of the image caption task based on CUB-200 dataset.

Text Promotion: this bird has wings that are brown and has a yellow bill .

Image Samples:

Text Promotion: a completely black bird, small body and head, with a black beak.

Image Samples:

Text Promotion: this particular bird has a belly that is tan and white .

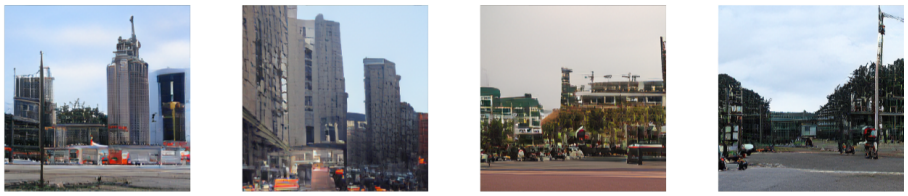
Image Samples:

Text Promotion: a parade of motorcycles is going through a group of tall trees .

Image Samples:

Text Promotion: a city street with multiple bildings and a street light .

Image Samples:



Text Promotion: an all white kitchen with an electric stovetop .

Image Samples:

Figure 8: More samples for the pure text-to-image generation from CUB-200 and MSCOCO.

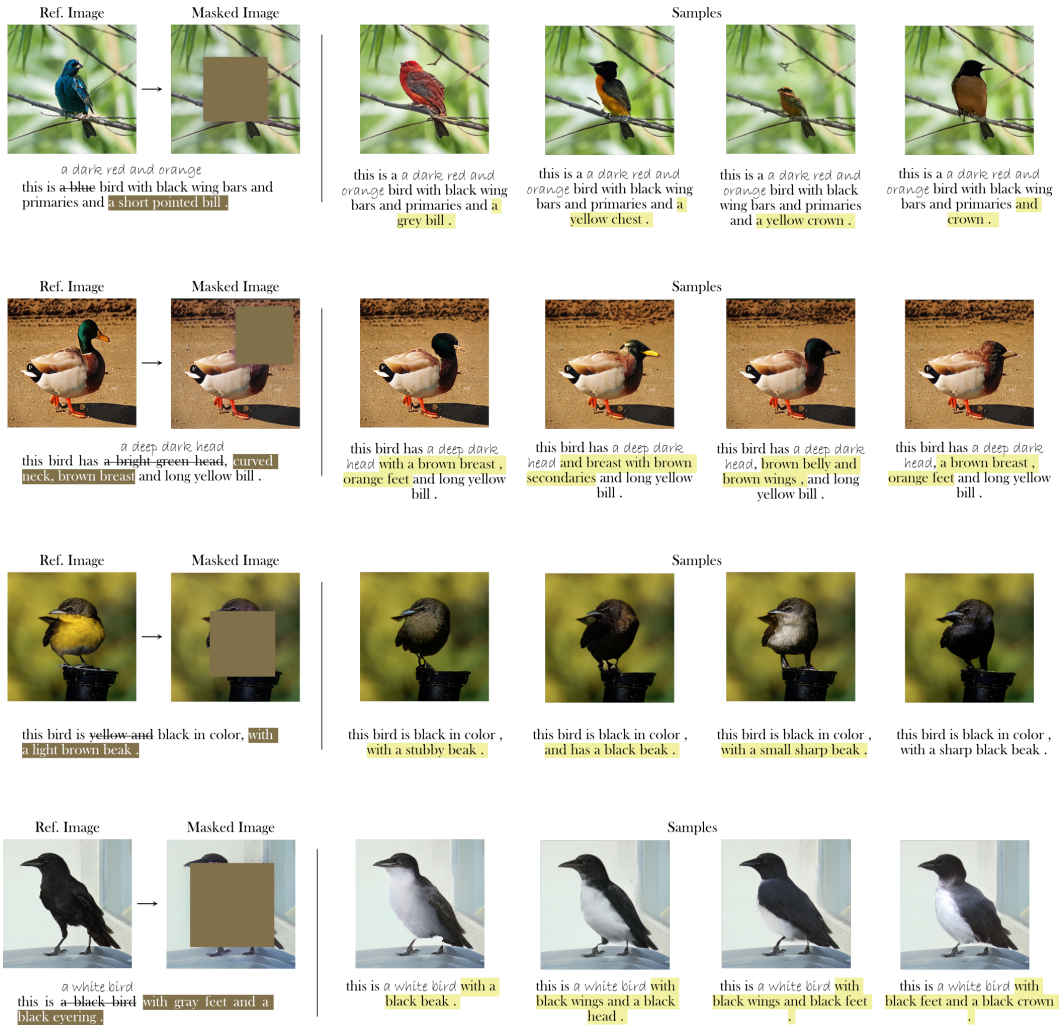



Figure 9: More samples of cross modal vision and language manipulation and infilling. The dark brown portions of the image and description represent the [MASK], while the strike-through represent the caption manipulation.

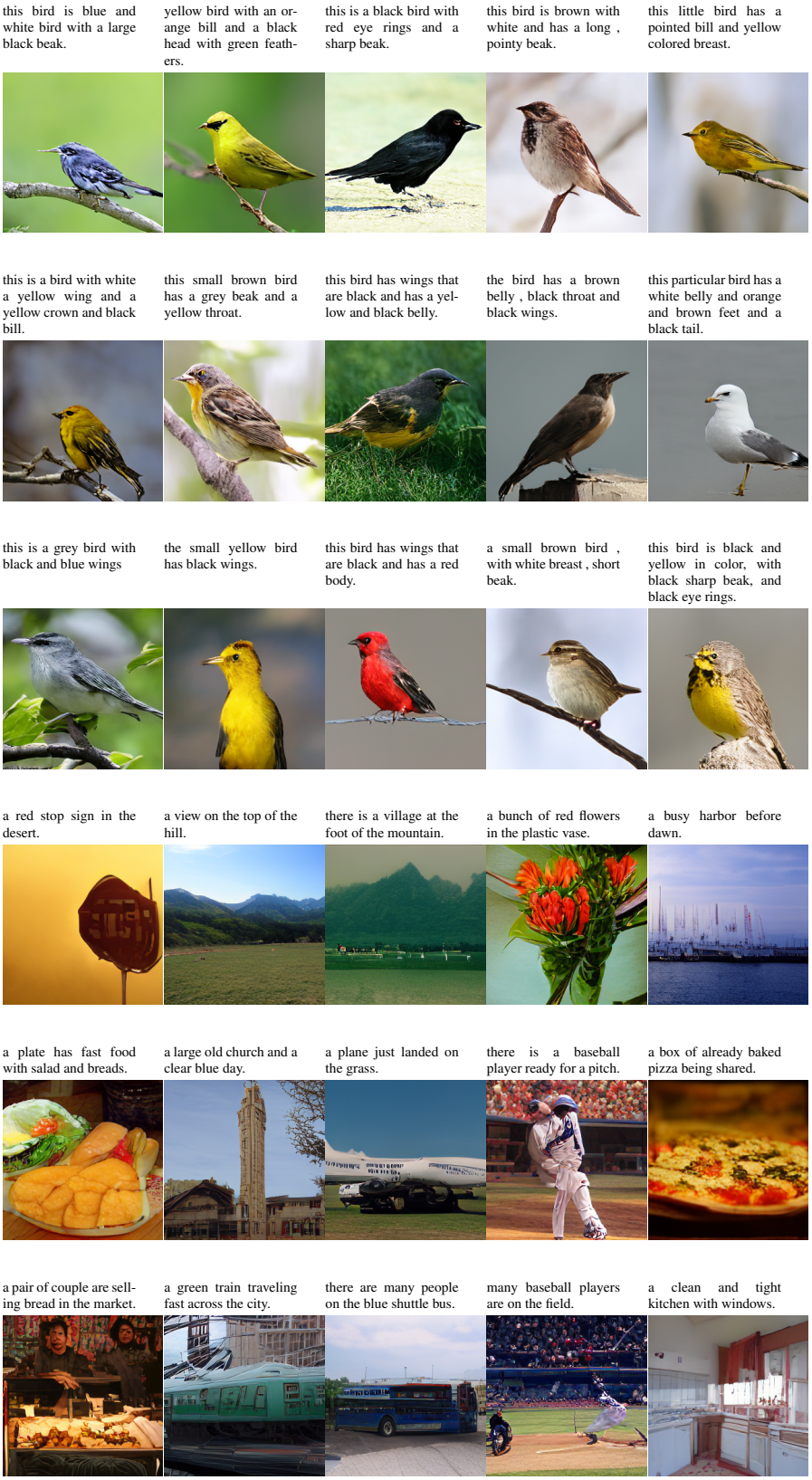



Figure 10: More vision-language pair samples from CUB-200 and MSCOCO. Both the image and caption are *generated simultaneously.*