Unconditional Image-Text Pair Generation with Multimodal Cross Quantizer

Hyungyung Lee[1]
ttumyche@kaist.ac.kr
Sungjin Park[1]
zxznm@kaist.ac.kr
Joonseok Lee[2, 3]
joonseok2010@gmail.com
Edward Choi[1]
edwardchoi@kaist.ac.kr

[1] KAIST
[2] Google Research
[3] Seoul National University

Abstract

Although deep generative models have gained a lot of attention, most of the existing works are designed for unimodal generation. In this paper, we explore a new method for unconditional image-text pair generation. We design Multimodal Cross-Quantization VAE (MXQ-VAE), a novel vector quantizer for joint image-text representations, with which we discover that a joint image-text representation space is effective for semantically consistent image-text pair generation. To learn a multimodal semantic correlation in a quantized space, we combine VQ-VAE with a Transformer encoder and apply an input masking strategy. Specifically, MXQ-VAE accepts a masked image-text pair as input and learns a quantized joint representation space, so that the input can be converted to a unified code sequence, then we perform unconditional image-text pair generation with the code sequence. Extensive experiments show the correlation between the quantized joint space and the multimodal generation capability on synthetic and real-world datasets. In addition, we demonstrate the superiority of our approach in these two aspects over several baselines. The source code is publicly available at: https://github.com/ttumyche/MXQ-VAE.

1 Introduction

Deep generative models focus mainly on unimodal generation, either unconditional (GAN [8], VAE [11], GPT [1]) or conditional (VQGAN [6], DALL-E [20]). Despite these influential works, studies on multimodal generation are still uncharted. One previous work [22] proposed generating image and text at the same time with a GAN-based approach. However, the core idea was to treat the text as an image, where the model generates two images, one for the image and another for the text. Thus, this process must undergo the OCR process [24].

In this paper, we design Multimodal Cross-Quantization VAE (MXQ-VAE), a novel vector quantizer that learns image-text representations to jointly generate image-text pairs without any conditional input and post-processing (*e.g.*, OCR), with which we discover that a joint representation space is effective for semantically consistent image-text pair generation.

To improve a multimodal semantic correlation in a quantized space, we combine VQ-VAE [16] with a Transformer encoder [26] and further apply an input masking. MXQ-VAE learns to discreteize masked image-text pairs into a quantized joint representation space and reconstruct them. Specifically, the Transformer encoder learns joint representations by performing a multi-head attention across the input, thereby can capture the semantic correlation between image and text. The input masking further enhances the correlation by making the masked part refer to the other modality to reconstruct the original input. Thus, we can convert the input to a unified code sequence, then train Autoregressive Transformer [18] to model a joint distribution over the sequence, allowing semantically consistent image-text pair generation.

We evaluate MXQ-VAE on one synthetic text-augmented MNIST, called Caption MNIST and three public benchmarks: Oxford Flower-102 [15], CUB-200-2011 [27], and COCO [13]. We observe that MXQ-VAE generates semantically consistent image-text pairs better than several baselines. Specifically, our approach achieves the highest average scores of 99.2% on Caption MNIST, outperforming the second highest baseline by + 4.7% and also improves the performance by + 0.8% on Flower, + 5.3% on CUB and + 2.3% on COCO.

In addition, to study the effectiveness of the quantized joint space for generating semantically consistent image-text pairs, we construct a corrupted dataset, called Degree dataset, by gradually adjusting the degree of alignment between image and text. The experimental result demonstrates that our approach can uphold the semantic correlation between image and text, while baselines fail. Furthermore, we show that this result leads to semantically consistent image-text pair generation.

Contributions of this paper can be summarized as follows:

- **Unconditional Image-Text Pair Generation**: We propose for the first time a novel vector quantization method, MXQ-VAE, that learns the quantized joint representation space for unconditional image-text pair generation.
- **Semantic Consistency of the Generated Samples**: Our experimental results reveal that MXQ-VAE generates a semantically consistent image-text pairs on multiple benchmark datasets, including Caption MNIST, Oxford Flower-102, CUB-200-2011, and COCO against several baselines.
- **Multimodal Semantic Correlation**: Additionally, the experimental results on the Degree dataset demonstrate that MXQ-VAE learns the meaningful semantic correlation between image and text in the quantized joint space. Furthermore, it turns out that the quantized joint space leads to semantically consistent image-text pair generation.

2 Related works

Generative Models Most generative models mainly focus on unimodal generation. VAE [11], GAN [8], and GPT [1] generate image or text without any conditional input. Recently, these studies have been prominent approaches for text-conditional image [6, 20] and image-conditional text generation [9, 12]. With these model variants, unimodal generation has been rapidly improved. However, studies on multimodal generation are still unexplored. Joint GAN [22] proposes unconditional image-text generation, but this model generates two images, one for the image and another for the text, thus it must undergo the OCR process [24]. Similarly, MMVAE [21] generates two images with a shared space on MNIST-SVHN [21]. However, both images depict simple digits. They also experiment with an image-text pair dataset. It generates text, but for the image, it retrieves the nearest-neighbor original image in the feature space. In this paper, we use MMVAE with customized decoder for comparison.

Vector Quantized Variational Autoencoder VQ-VAE [16] is a representative model that maps continuous input into discrete representations by adopting an encoder-decoder architecture with a fixed size learnable codebook. In this paper, we aim to jointly generate image-text pairs without any conditional input. To achieve this, our model requires a continuous or discrete joint representation space, since image is inherently continuous and text is discrete. While a continuous space is a predominant approach for joint representation learning, we adopt a discrete space for the following reasons. First, powerful autoregressive generative model [1] has been developed to model distributions over discrete variables. Next, the discrete space does not suffer from several drawbacks common to the continuous space, such as posterior collapse in VAE [11]. In addition, discrete variables have the advantage of being more interpretable and space-efficient than continuous variables [2]. Consequently, we propose a simple approach based on VQ-VAE, which learns a quantized joint representation space.

Joint Image-Text Representations Following the success of Transformer [26] in NLP tasks, there is a simultaneous explosion of Transformer-based models in joint representation learning. Previous works (e.g. UNITER [3], Pixel-BERT [10], VLP [29]) utilize BERT [4], a Transformer encoder-based model, to learn joint image-text representations. By leveraging the bidirectional self-attention mechanism of BERT, both images and text can capture the semantic correlation between them without requiring annotations that align image and text.

3 Multimodal Cross-Quantization VAE (MXQ-VAE)

Our goal is to generate semantically consistent image-text pairs simultaneously without any conditional input. To achieve this, we learn a joint representation space by quantizing both image and text into a discrete space based on VQ-VAE. As shown in Fig. 1, we adopt a two-stage approach.

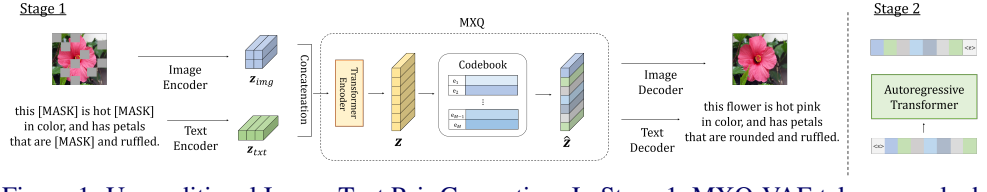

Figure 1: Unconditional Image-Text Pair Generation. In Stage 1, MXQ-VAE takes a masked image-text pair as input, and learns a quantized joint representation space. Then, the input is converted into a unified code sequence. In Stage 2, Autoregressive Transformer models a joint distribution over the code sequence. At inference, MXQ-VAE decodes a sampled code sequence to an image-text pair.

3.1 Stage 1: Learning a Quantized Joint Representation Space

MXQ-VAE learns to discretize image-text pair into a quantized joint representation space and reconstructs them. It consists of three major parts: Encoders, Decoders, and MXQ. The MXQ module contains the Transformer encoder and a codebook $C = \{e_m\}_{m=1}^{M}$ of size M, where $e_m \in \mathbb{R}^d$. Each encoder and decoder is 2D CNN for image and 1D CNN for text. The effect of other architectural choices is discussed in the experiments.

Input Masking Given an image-text pair, we first split the image $I \in \mathbb{R}^{H \times W \times 3}$ into non-overlapping patches of equal size and the text into tokens $T = \{t_1, ..., t_N\}$ with WordPiece

[28]. Then, we randomly mask the patches and the tokens with a probability p from a uniform distribution without replacement. Each pixel in the masked patches is set as zero. We replace the masked tokens with a special token `[MASK]`. This approach makes two modalities complementary. To reconstruct the masked part, the model should not only refer to the unmasked part of each modality (*i.e.*, intra-modal), but also to the unmasked part of the other modality (*i.e.*, cross-modal). In this way, the model can learn the relationship between image and text.

Encoders A masked input image is encoded to a set of image feature map $\mathbf{z}_{img} \in \mathbb{R}^{h\times w\times d}$. Similarly, a masked text input is encoded to $\mathbf{z}_{txt} \in \mathbb{R}^{n\times d}$, where both of them are downsampled by a factor of f; that is, $h = \frac{H}{f}, w = \frac{W}{f}, n = \frac{N}{f}$.

MXQ The Transformer encoder takes the concatenation of $\mathbf{z}_{img}$ and $\mathbf{z}_{txt}$ as input, and produces joint image-text representations $\mathbf{z} \in \mathbb{R}^{\ell\times d}$, where $\ell = h\times w + n$. The output is discretized into the quantized joint space by performing the nearest-neighbor search in the codebook C as given in Eq. (1) and produces a unified code sequence $\hat{\mathbf{z}} \in \mathbb{R}^{\ell\times d}$. With this simple approach, a discrete code can contain the correlated information of image and text.

$$\hat{\mathbf{z}}_i = \text{Quantize}(\mathbf{z}_i) = e_m \quad \text{where } m = \underset{j}{\text{argmin}} \|\mathbf{z}_i - e_j\| \tag{1}$$

Decoders We first apply a linear layer to the spatial dimension (*i.e.*, ℓ) of $\hat{\mathbf{z}}$ to ensure that the decoder takes the desired size as input and produces $\hat{\mathbf{z}}_{img}$ and $\hat{\mathbf{z}}_{txt}$ for image and text, respectively. The decoder then reconstructs the original input from $\hat{\mathbf{z}}_{img}$ and $\hat{\mathbf{z}}_{txt}$, yielding reconstruction results, $I' \in \mathbb{R}^{H\times W\times 3}$ and T', respectively.

Our model is optimized using the following objective:

$$L = \delta_1 \underbrace{\|I - I'\|_2^2}_{\text{image recon loss}} - \delta_2 \underbrace{\log p(T|\hat{\mathbf{z}}_{txt})}_{\text{text recon loss}} + \delta_3 \underbrace{\|sg[\mathbf{z}] - \hat{\mathbf{z}}\|_2^2}_{\text{codebook loss}} + \delta_4 \underbrace{\|sg[\hat{\mathbf{z}}] - \mathbf{z}\|_2^2}_{\text{commitment loss}} \tag{2}$$

where each loss term is weighted by δ_i and sg refers to a stop-gradient.

3.2 Stage 2: Unconditional Image-Text Pair Generation

We adopt the Autoregressive Transformer [18] architecture to model a joint distribution over the sequence of unified code indices $c = (c_1, c_2, ..., c_\ell)$ from Stage 1. The probability of each code index in the sequence is conditioned on all previously predicted code indices $c_{<n} = (c_1, c_2, ..., c_{n-1})$ and the joint distribution of the sequence is obtained as the product of conditional distributions: $p(c) = \prod_{n=1}^{\ell} p(c_n|c_1, c_2, ..., c_{n-1}) = \prod_{n=1}^{\ell} p(c_n|c_{<n})$. During training, MXQ-VAE quantizes the input image-text pair into the unified code sequence, then Autoregressive Transformer is trained to predict the next code index in the given sequence. At inference, we sample a code sequence from Autogregressive Transformer via Top-k sampling [7], then MXQ-VAE decodes the sampled code sequence to an image-text pair.

4 Experimental Settings

4.1 Datasets

Caption MNIST. Following [23], we build 600k synthetic image-text pairs. Each pair contains several colors, digits, and positions. We have 4 colors (white, red, green, and blue), 10 digits (0 to 9) and 5 positions (center, top left, top right, bottom left, and bottom

right). According to the filled quadrant, we refer to each subset as Single and Quad1 to Quad4. For example, Single pairs only have one colored digit at the center of the image and a corresponding caption, whereas Quad3 pairs have colored digits in three quadrants, also with a corresponding caption. See Fig. 2 for more details.

Oxford Flower-102 [15] (Flower) contains 8,189 flower images with 10 captions per image.

CUB-200-2011 [27] (CUB) consists of 11,788 bird images with 10 captions per image. We use a bounding box to cut the background of the image and only use the content.

COCO [13] is a real-world dataset with about 120k images and 5 captions per image.

Degree datasets. To evaluate the semantic correlation between image and text in the quantized joint space in Stage 1, we construct the Degree dataset by gradually adjusting the degree of alignment between image and text. More specifically, for Caption MNIST, we replace the color and digit in the caption with other random colors and digits. For instance, Quad3 can have 4 degrees from perfectly paired (Degree 3) to completely unpaired (Degree 0). Refer to Fig. 2 for more details. For Flower and CUB, variables besides color are difficult to control, thus we only consider the number of unique colors in the caption and replace them with other random colors. According to the number of unique colors in the caption, we refer it as Quad1 to Quad4. See Fig. 3 for more details.

Quadrant	Image	Text
Single		This is white 7.
Quad1		The 6 on the lower right is blue.
Quad2		The green 0 is on the upper right, and the top left 8 is red.

Quadrant	Image	Text
Quad3		the lower left 4 is blue, the red 0 is on the lower right, and the 2 on the top right is green.
Quad4		The bottom left 1 is blue, the upper right 7 is white, the bottom right 0 is red, and the upper left 3 is green.

Image	Degree	Text
	3	the lower left 4 is blue, the red 0 is on the lower right, and the 2 on the top right is green.
	2	the lower left 5 is white, the red 0 is on the lower right, and the 2 on the top right is green.
	1	the lower left 1 is red, the green 7 is on the lower right, and the 2 on the top right is green.
	0	the lower left 8 is white , the blue 2 is on the lower right, and the 9 on the top right is red.

Figure 2: Examples of Caption MNIST (Top) and the Quad3 Degree dataset (Bottom).

Image	Degree	Text
	2	this flower has petals that are yellow and has brown stamen.
	1	this flower has petals that are red and has brown stamen.
	0	this flower has petals that are maroon and has wine stamen.

Image	Degree	Text
	2	this is a small bird with a bright yellow breast and a black cheek patch.
	1	this is a small bird with a bright orange breast and a black cheek patch.
	0	this is a small bird with a bright purple breast and a pink cheek patch.

Figure 3: Examples of the Flower (Left) and CUB (Right) Quad2 Degree datasets.

4.2 Evaluation Metrics

4.2.1 Multimodal Semantic Correlation Evaluation

We evaluate the semantic correlation between image and text in the quantized joint space by the text reconstruction accuracy on the Degree dataset. Since the quantized joint space captures the semantically correlated parts of images and text, the model should identify the corrupted parts in the Degree dataset and not reconstruct that part *as is*. For instance, in the Caption MNIST Quad1 Degree dataset, the Degree 1 input text is “the green 0 is on the upper right.” and the Degree 0 input text is “the white 1 is on the upper right.”. The input image depicts the Degree 1 text in this case. If the reconstructed text is “the green 0 is on the upper right.” for both Degree 1 and 0 input text, the text reconstruction accuracy will be 1 and 0 for

each, as we only consider color and digit for calculating text reconstruction accuracy (only color in cases of Flower and CUB). According to this, the desired accuracy, for instance, would be 1, 0.67, 0.33, and 0 for each degree of the Quad3 Degree dataset.

4.2.2 Generated Image-Text Alignment Evaluation

We evaluate the semantic consistency of the generated image-text pairs with rule-based semantic parser on Caption MNIST, label-based modified unigram precision [17] and sentence similarity on Flower and CUB, and CLIP-based retrieval on COCO.

Rule-based Semantic Parser. Following [23], we extract a set of position, color and digit of the generated text with the rule-based parser. With a color and digit classifier trained on Caption MNIST images that achieved 100% and 99.5% accuracy respectively, we predict the color and digit of the position that corresponds to the parsed text in the generated image and measure whether both the predicted color and digit match the parsed text at that position.

Label-based Evaluations. For Flower and CUB, we measure the semantic consistency between the generated caption and all original captions that belong to the same label as the generated image. Specifically, we first train image classifiers using the original image and achieve 99% and 93% accuracy for Flower and CUB, respectively. Then, we predict the label of the generated image and collect all original captions from the same label. For the modified unigram precision [17], we report the average score of the multiset intersection of words in the original text and the generated text divided by the total number of words in the generated text. For the sentence similarity, a pre-trained BERT [4] takes the generated text and the original text from the same label, separately, and outputs a mean-pooled vector. Then, we calculate the cosine similarity between them and report a Top-1, 5 and 10 averaged score.

CLIP-based Retreival. For COCO, we report the Precision@{1, 5, 10} to measure the retrieval accuracy of CLIP [19] of the generated text from the 100 text candidates; that is, 1 positive from the generated text, 99 random negatives from the original text.

4.3 Baselines

MXQ-VAE w/o IM (Input Masking). The architecture is the same as MXQ-VAE, but without input masking.

MXQ-VAE w/o TC (Text Compression). This replaces the 1D convolution-based text encoder and decoder with the Transformer encoder. Note that the Transformer text encoder outputs the same number of embeddings as the input, contrary to 1D convolutional layers that compresses the input text gradually by each layer. This is why we name it *w/o **T**ext **C**ompression*.

Unimodal Quantizer. In Stage 1, this baseline discretizes each modality separately. For the image, it follows the original VQ-VAE. Since the text is originally discrete, we directly use the word embeddings. In Stage 2, we concatenate the code sequence of image and text embeddings as the input. Depending on which modality comes first in Stage 2 input, we refer to it as I_$\mathrm{T}_{\mathrm{Embd}}$ or $\mathrm{T}_{\mathrm{Embd}}$_I.

Only Sharing *C*. In Stage 1, this baseline only shares the codebook C without the Transformer module that combines image and text together and the input masking. In Stage 2, depending on which modality comes first, we refer to it as I&T or T&I.

The variants of Unimodal Quantizer and Only Sharing C also generate an image-text pair without any conditional input, and the only difference is which modality comes first in Stage 2.

4.4 Implementation Details

Stage 1. For Caption MNIST, the codebook C is 256×128, the input size is $64 \times 64 \times 3$ image and 64×128 word embedding. Each input is downsampled by a factor of $f = 8$. We adopt 2 stacks of the Transformer encoder and apply an input masking ratio $p = 0.3$. We use a batch size of 800 with a learning rate of 5×10^{-4}. We set δ_1, δ_2 and δ_3 to 1.0 and δ_4 to 0.25. **Stage 2.** We adopt GPT-2 [18] for an autoregressive generative model with 8 layers, 8 attention heads and 512 embedding dimensions. We adopt Top-k sampling [7] with $k = 10$. For Caption MNIST, we use a batch size of 800 with a learning rate of $5e-4$. In all our experiments, we use AdamW [14] with $\beta_1 = 0.9$, $\beta_2 = 0.99$ with cosine decay learning rate scheduler and train the model using a NVIDIA RTX A6000.

5 Results and Discussion

5.1 Multimodal Semantic Correlation Results

We first study **the effectiveness of MXQ-VAE in constructing multimodal semantic correlation in the quantized joint space** as described in Sec. 4.2.1. Tab. 1 shows the results. On the Caption MNIST Degree dataset, we observe that Only Sharing C and MXQ-VAE w/o TC cannot fully capture the semantic correlation between image and text. In fact, MXQ-VAE w/o TC completely fails to learn the relationship between image and text. Our approach, on the other hand, shows the best approximation in all quadrants. This result suggests that MXQ-VAE can identify the correlation and the difference between image and text. Also, we see the advantage of the input masking that brings in considerable improvement. Moreover, our approach again achieves the best performance over baselines on the Flower and CUB Degree datasets. See more results in supplementary. The above results show that the Transformer encoder for cross-modal attention and the text compression are essential for the multimodal semantic correlation, and the input masking also plays a significant role. Consequently, we choose MXQ-VAE as our final design for unconditional image-text pair generation in Stage 2.

We also visualize the unified code sequences and the attention maps of the Transformer encoder in the MXQ module. Fig. 4 (a) and (b) show the t-SNE [25] visualizations of the unified code sequences with MXQ-VAE and Only sharing C for 10 digits per color on Caption MNIST Single image-text pairs. MXQ-VAE has a unique cluster for each digit, while the baseline has two. This is because there are two types of text in the Single pairs: 1) This {digit} is {color}; 2) This is {color} {digit}. As shown in Fig. 4 (c) and (d), MXQ-VAE can capture the correlation between them, but the baseline completely fails even though they contain the same content. Fig. 5 visualizes the attention maps on the text tokens when the image patch is given as a query. These results again show the superiority of MXQ-VAE.

5.2 Generated Image-Text Alignment Results

We evaluate **the semantic consistency of the generated image-text pairs** as described in Sec. 4.2.2. Tab. 2 shows the results on Caption MNIST. We can observe that MXQ-VAE outperforms all baselines on every quadrant. Also, note that all baselines are vulnerable to which modality is given first to generate the image-text pairs. All models given text first significantly underperform up to 21.7% (in I_T_{Embd} and T_{Embd}_I) on average compared to the model given image first. We assume that this is due to the fact that quantized image is longer than text and image often contains more complex (and complete) information than text. MXQ-VAE, on the other hand, avoids this problem with a unified code sequence. We

Dataset	Degree	Models	Degree 4	Degree 3	Degree 2	Degree 1	Degree 0
Caption MNIST	Quad4	Only Sharing C	0.486	0.443	0.394	0.358	0.315
		MXQ-VAE w/o TC	1.0	0.975	0.951	0.929	0.906
		MXQ-VAE w/o IM	0.896	0.802	0.698	0.595	0.498
		MXQ-VAE (Ours)	0.969	0.729	0.489	0.248	0.009
	Quad3	Only Sharing C	-	0.713	0.64	0.558	0.483
		MXQ-VAE w/o TC	-	1.0	0.968	0.943	0.909
		MXQ-VAE w/o IM	-	0.999	0.862	0.714	0.564
		MXQ-VAE (Ours)	-	0.997	0.663	0.334	0.012
Flower	Quad4	Only Sharing C	0.939	0.704	0.516	0.321	0.131
		MXQ-VAE w/o TC	1.0	0.944	0.886	0.866	0.810
		MXQ-VAE w/o IM	0.997	0.728	0.482	0.278	0.067
		MXQ-VAE (Ours)	0.996	0.737	0.490	0.250	0.014
	Quad3	Only Sharing C	-	0.959	0.675	0.426	0.158
		MXQ-VAE w/o TC	-	1.0	0.927	0.870	0.816
		MXQ-VAE w/o IM	-	0.997	0.662	0.377	0.090
		MXQ-VAE (Ours)	-	0.999	0.660	0.339	0.019
CUB	Quad4	Only Sharing C	0.985	0.771	0.572	0.356	0.155
		MXQ-VAE w/o TC	1.0	0.948	0.894	0.825	0.748
		MXQ-VAE w/o IM	0.998	0.833	0.645	0.424	0.181
		MXQ-VAE (Ours)	0.995	0.749	0.515	0.292	0.083
	Quad3	Only Sharing C	-	0.986	0.755	0.498	0.199
		MXQ-VAE w/o TC	-	1.0	0.938	0.864	0.772
		MXQ-VAE w/o IM	-	0.998	0.806	0.559	0.248
		MXQ-VAE (Ours)	-	0.995	0.709	0.421	0.119

Table 1: Multimodal semantic correlation on the Caption MNIST, Flower and CUB Degree datasets. The scores close to 1.0, 0.75, 0.5, 0.25, 0.0 on Quad4 and 1.0, 0.67, 0.33, 0.0 on Quad3 are better. In general, our approach shows the best performance.

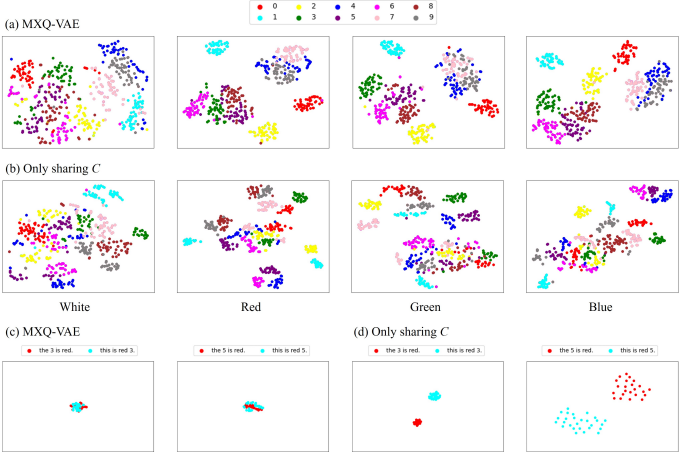

Figure 4: t-SNE visualizations of the unified code sequence on Caption MNIST Single image-text pairs. In (a) and (b), MXQ-VAE has a unique cluster for each digit compared to the baseline. In (c) and (d), unlike MXQ-VAE, the baseline cannot identify the correlation between the two types of text, even though they contain the same content.

report the results on Flower and CUB in Tab. 3. Compared to the baselines, MXQ-VAE performs well in all metrics, indicating its capability to generate semantically consistent image-text pairs for real-world data as well as carefully controlled synthetic data. Also, we again demonstrate the superiority of MXQ-VAE on COCO in Tab. 4. This result indicates that our approach can be extended to large-scale data in the future.


a flower with long and wide petals that are purple with yellow center .


a flower with very light pink petals with a yellow center .


this small bird has a yellow head and breast with a white belly .


a big bird with a white under ##bell ##y and gray and black feathers .

Figure 5: Visualization of the attention maps on Flower and CUB. The deeper color indicates a higher attention score. Although we do not provide any explicit guidance, MXQ-VAE can properly attend to the related tokens that correspond to the image patch.

Fig. 6 shows the comparison of generated image-text pairs on Caption MNIST. We can see that MXQ-VAE can generate semantically consistent image-text pairs and also show high-fidelity image and text. In contrast, although I_T_{Embd} can generate high-quality text, the generated images are partially blurry and hard to identify the digit. Also, I&T can generate high fidelity images, but the generated pairs are irrelevant, showing its poor ability to capture the relationship between image and text. Similarly, MXQ-VAE shows its superiority on Flower and CUB as shown in Fig. 7. Unlike MXQ-VAE, some parts of the images of I_T_{Embd} are incomplete. Also, I&T and MMVAE [21] cannot generate detailed parts, and most of the generated images of bird are blurry, and they generate overlapping words. In addition, Joint GAN [22] fails to generate realistic images and the generated text is grammatically incorrect. The underline in the figure represents the mismatched part between image and text. For more samples, refer to supplementary.

Models	Single	Quad1	Quad2	Quad3	Quad4	Average
I&T	0.979	0.926	0.675	0.434	0.255	0.654
T&I	0.803	0.780	0.458	0.282	0.161	0.497
I_T_{Embd}	0.953	0.953	0.956	0.958	0.849	0.945
T_{Embd}_I	0.086	0.895	0.913	0.916	0.828	0.728
MXQ-VAE (Ours)	**0.998**	**0.997**	**0.994**	**0.996**	**0.974**	**0.992**

Table 2: Semantic consistency of the generated image-text pairs on Caption MNIST

In summary, extensive results demonstrate that the quantized joint space with multimodal semantic correlation is effective for semantically consistent image-text pair generation. Furthermore, our approach enhances the meaningful semantic correlation between image and text and also encourages the generation of semantically consistent image-text pairs.

Models	Flower				CUB			
	Modified unigram	Sentence similarity			Modified unigram	Sentence similarity		
	precision	Top-1	Top-5	Top-10	precision	Top-1	Top-5	Top-10
I&T	0.420	0.935	0.920	0.910	0.418	0.912	0.889	0.875
T&I	0.402	0.910	0.891	0.880	0.425	0.917	0.896	0.884
I_T_{Embd}	0.406	0.938	0.923	0.913	0.425	0.921	0.900	0.886
T_{Embd}_I	0.392	0.927	0.900	0.884	0.423	0.917	0.895	0.882
Joint GAN [22] *	0.324	0.808	0.788	0.774	-	-	-	-
MMVAE [21]	-	-	-	-	0.262	0.707	0.682	0.667
MXQ-VAE (Ours)	**0.428**	**0.941**	**0.926**	**0.916**	**0.478**	**0.948**	**0.919**	**0.900**

Table 3: Semantic consistency of the generated image-text pairs on Flower and CUB. * means that we measure the score with the image-text pairs reported in Joint GAN [22].

Models	P@1	P@5	P@10
I&T	0.062	0.263	0.328
I_T_{Embd}	0.083	0.29	0.425
MXQ-VAE (Ours)	**0.106**	**0.323**	**0.491**

Table 4: Semantic consistency of the generated image-text pairs on COCO

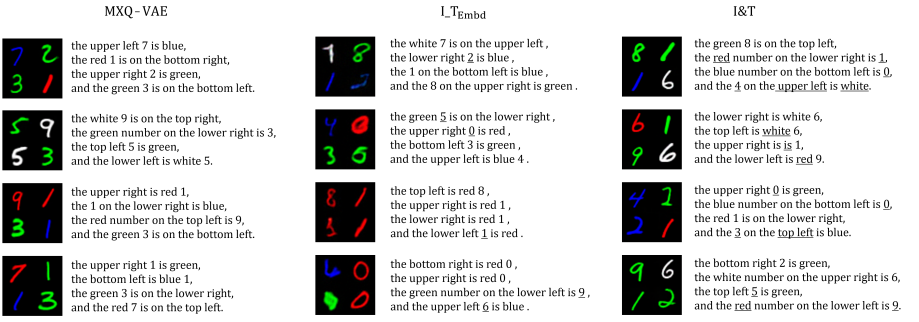

Figure 6: Comparison of generated image-text pairs on Caption MNIST

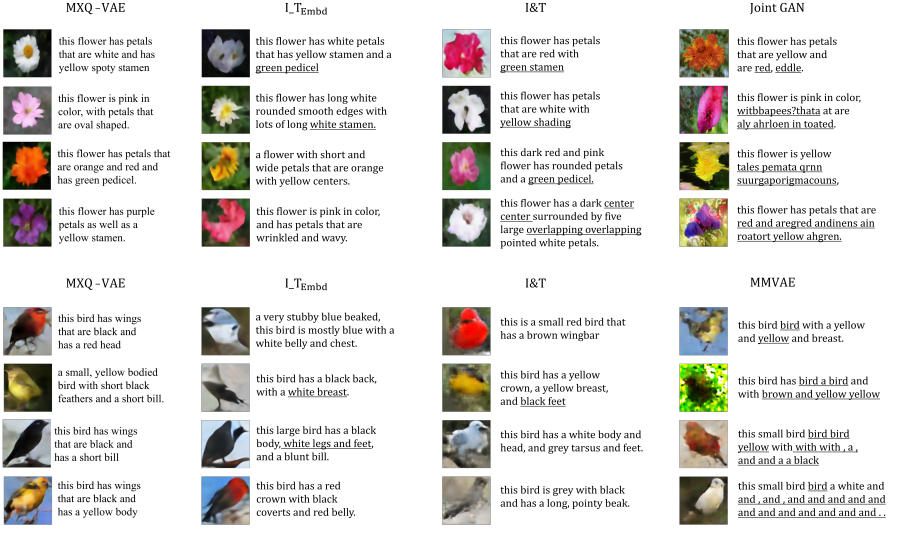

Figure 7: Comparison of generated image-text pairs on Flower and CUB

6 Conclusion and Future Work

In this study, we propose for the first time a novel vector quantizer, MXQ-VAE, that learns a quantized joint representations space for unconditional image-text pair generation. We demonstrate that the quantized joint space with a multimodal semantic correlation encourages the generation of semantically consistent image-text pairs. With extensive experiments, our approach shows superiority in these two aspects over several baselines. We hope our work suggests new directions for multimodal generation and joint representation learning. Besides exploring AI's multimodal creativity, one promising application of our approach is training conditional generative models with three or more modalities (*e.g.*, image, text, audio). By considering MXQ-VAE as a multimodal tokenizer, generative models can generate other modalities given a multimodal code sequence (or vice versa), which remains as future work.

Acknowledgements

This work was supported by the KAIST-NAVER Hyper-Creative AI Center and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.2019-0-00075 Artificial Intelligence Graduate School Program(KAIST) and No.2021-0-01778 Development of human image synthesis and discrimination technology below the perceptual threshold), and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945 and NRF-2021H1D3A2A03038607) funded by the Korea government (MSIT).

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv:2005.14165*, 2020.

[2] Ruben Cartuyvels, Graham Spinks, and Marie-Francine Moens. Discrete and continuous representations and processing in deep learning: looking forward. *AI Open*, 2:143–159, 2021.

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

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[5] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835, 2021.

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[7] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv:1805.04833*, 2018.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[9] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022.

[10] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv:2004.00849*, 2020.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

[12] 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*, pages 121–137. Springer, 2020.

[13] 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*, pages 740–755. Springer, 2014.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.

[15] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[16] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv:1711.00937*, 2017.

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

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

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

[20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv:2102.12092*, 2021.

[21] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Brian Shimanuki. *Joint generation of image and text with GANs*. PhD thesis, Massachusetts Institute of Technology, 2019.

[23] Woncheol Shin, Gyubok Lee, Jiyoung Lee, Joonseok Lee, and Edward Choi. Translation-equivariant image quantizer for bi-directional image-text generation. *arXiv:2112.00384*, 2021.

[24] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR)*, volume 2, pages 629–633. IEEE, 2007.

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[26] 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.

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

[28] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*, 2016.

[29] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.