Augmenting Vision Language Pretraining by Learning Codebook with Visual Semantics

*Xiaoyuan Guo[†], *Jiali Duan[‡], C.-C. Jay Kuo[‡], Judy Wawira Gichoya[§] and Imon Banerjee[¶,‖]
[†]Department of Computer Science, Emory University, GA, USA
[‡]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, CA, USA
[§]Department of Radiology and Imaging Sciences, Emory University, GA, USA
[¶]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, AZ, USA
[‖]Department of Radiology, Mayo Clinic, AZ, USA
Email: xiaoyuan.guo@emory.edu, jialidua@usc.edu, cckuo@sipi.usc.edu, judywawira@emory.edu, banerjee.imon@mayo.edu

***Abstract*—Language modality within the vision language pretraining framework is innately discretized, endowing each word in the language vocabulary a semantic meaning. In contrast, visual modality is inherently continuous and high-dimensional, which potentially prohibits the alignment as well as fusion between vision and language modalities. We therefore propose to "discretize" the visual representation by joint learning a codebook that imbues each visual token a semantic. We then utilize these discretized visual semantics as self-supervised ground-truths for building our Masked Image Modeling objective, a counterpart of Masked Language Modeling which proves successful for language models. To optimize the codebook, we extend the formulation of VQ-VAE which gives a theoretic guarantee. Experiments validate the effectiveness of our approach across common vision-language benchmarks.**

I. Introduction

Inspired by the success of language modeling [1], [2], the concept of Vision-Language Pretraining (V&L) has attracted growing attention in the community, where the model is pretrained once and achieves superior performance over a set of downstream tasks, via transfer learning. The central theme to the problem is learning better alignment and interactions between the two modalities via feature fusion. Two popular ways - late fusion and early fusion have been researched. Late fusion approaches such as CLIP [6] and ALIGN [7] directly optimize the InfoNCE [8] objective, by leveraging large amount of paired data (400M for CLIP and 1.8B for ALIGN). On the other hand, early fusion approaches such as VinVL [3], ViLT [4] and OSCAR [5] adopt a multi-modal transformer to model the interactions between the vision and text modalities. On top of which, objectives such as image-text matching (ITM) and masked language modeling (MLM) are optimized to enforce the alignment. Our work belongs to this category.

MLM has proven successful for language modeling. The key idea is to predict masked text tokens given an image and unmasked text tokens. Questions naturally arise, *can we do the same for the image modality? Will it be as effective as MLM for performance?* The main hurdle for applying masked image modeling (MIM) lies in the difference between vision and language. Language is inherently discrete, endowing each word in the language vocabulary a semantic meaning. In contrast, visual modality is continuous and high-dimensional. Existing approaches such as VinVL [3], OSCAR [5], UNITER [9] utilize an object detector to assign a class ID to the visual patches, based on which the MIM objective is built. However, the number of classes that an object detector can recognize is limited and the detector is not end-to-end optimized.

To overcome the challenge, we propose a **C**ode**B**ook based approach for **Vi**sion **L**anguage **A**lignment (**CB-ViLA**), which can help quantize visual features and facilitate optimizing image-text alignment as well as the MIM objective. Inspired by the ability of learning discrete visual representations of VQ-VAE [10], we extend its formulation into the multi-modal setting to learn a codebook with visual semantics. Specifically, we approximate the latent space posterior $q(z|x)$ with a visual encoder and the codebook is then used to "discretize" the latent space into a probability distribution over the codebook vectors $q(z = c_k|x)$. During decoding, we incorporate the language modality into the VAE reconstruction objective by taking text sequence features as a conditional variable. In this way, our multi-modal fusion encoder can also be interpreted as VQ-VAE decoder, where its parameters are shared for multi-modal interactions and VQ-VAE decoding (See Figure 1).

Although codebook is also utilized in works such as BEiT [12], DALLE [13], they are different from our design. As the codebook used in BEiT [12] is off-the-shelf from DALLE [13], kept frozen during pretraining. While in SOHO [14], the codebook is heuristically updated via momentum. Instead, we integrate codebook learning into the vision-language pretraining framework by alternately optimizing the codebook and the encoders (Bert [2], ViT [15] and multi-modal fusion encoder). The codebook gradient is frozen when computing MIM while the encoders are updated.

To summarize, our main contributions are as follows:

1) We present a vision language framework that unifies MLM and MIM by jointly optimizing a visual semantic codebook, on the evidence lower bound of language-conditioned pixel reconstruction posterior.
2) We show that the quantization codebook is able to learn useful visual semantics, and together with the MIM objective, help improve the performance of the framework over downstream tasks.

* Guo and Duan contributed equally to the work.

II. Related Work

Vision-Language Pretraining. V&L has been popular recently as it enables transfer of superior performance in a wide variety of downstream vision-language tasks by pre-training once, similar to language modelling [1], [2], [9]. Existing works can be categorized into one-stream v.s. two-stream. In one stream model, features of different modalities are directly fed into a transformer [16]. Whereas in two stream models, inputs are first processed by two single-modal networks before fed into a transformer [17]. More recently, some works such as ALBEF [18], CODIS [44], TCL [22] adopt a hybrid architecture, which combines both one-stream muti-modal encoder with two-stream uni-modal encoders. Our architecture falls into this category. The main difference between ALBEF [18] and ours is visual semantic codebook learning and masked image modeling objective designed to address the discrepancy between image and text modalities.

Codebook Usage. Some recent works have adopted codebook for vision language pretraining. For example, CODIS [44] proposes a multi-modal codebook as a bridge to align image and text modalities at a cluster level. The multi-modal codebook is optimized with optimal transport. The codebook in SOHO [14] discretizes features from the convolutional encoder and is momentum updated. Although we also focus on learning a visual quantization codebook for multi-modal alignment, the learning process is different as our codebook is updated by optimizing the evidence lower bound of language-conditioned pixel reconstruction posterior. In this sense, our framework can also be seen as an extension of VQ-VAE [10] where ViT and multi-modal fusion are treated as encoder and decoder respectively. The idea of our codebook shares some similarity with the concept "vector quantization" of VideoBert [41], which is achieved via hierarchical k-means with human inspection. Other than the applications in V&L, codebook has be utilized for vision tasks [23], [12] as well. For example, BEiT [12], VIOLET [42] both use a pretrained codebook for input space tokenization. These visual codebooks are kept fixed during training. To further improve the quality of visual codebook, PeCo [43] extends DALLE [13] with a perceptual loss. Notably, our codebook is joint optimized and updated to tokenize the feature space.

III. Method

Figure 1 is an overview of our **CB-ViLA** framework. Our goal is to learn visual semantics via a codebook, facilitating the alignment between image and text modalities. We describe visual semantic codebook learning in Section III-A. In Section III-B, we explain how the codebook is integrated. In particular, the "discretized" visual information will be used as MIM targets and multi-modal queries respectively in a self-supervised manner. Finally, we illustrate how our proposed components fit into the vision-language pretraining framework as well as training procedure.

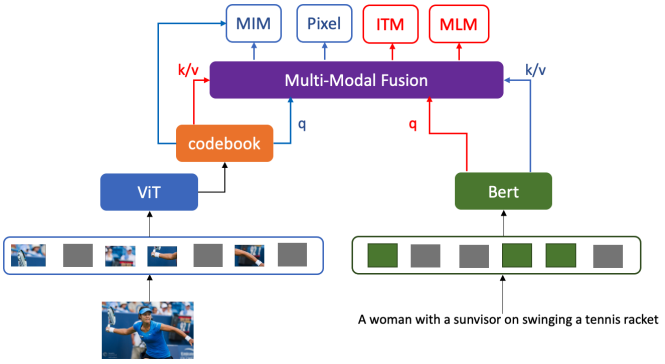

Fig. 1. Overview of **CB-ViLA** framework. We leverage a codebook to enforce visual semantic alignment, via the use of text-conditioned masked image modeling. Arrows of the same color indicate the flows for a specific pretraining objective. For example, MIM is calculated with codeword indices and the *cls* token of multi-modal fusion module, which is a cross-attention of "discretize" ViT [15] output as query and Bert [2] as key/value. Our codebook is optimized via an extension of VQ-VAE [10] objective.

A. Visual Semantic Codebook Learning

As mentioned in Section I, language is inherently discrete and semantic rich, whereas the visual signal is continuous and could be noisy. To facilitate the alignment between the two, we leverage a learnable codebook to assign a visual semantic to each visual token.

We denote the learnable codebook as $\mathbf{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\} \in \mathcal{R}^{d_c \times K}$, where d_c is the dimension for each code and K equals to the number of codewords (e.g., 3K). Each $\boldsymbol{c} \in \mathbf{C}$ corresponds to a visual semantic (Refer to Section IV-G for visualization). Given an encoded visual sequence $\{v^1, v^2, ..., v^N\}$, where N denotes the length of visual sequence, the codebook "discretizes" them by doing a nearest neighbor look-up using the shared embedding space $\mathcal{R}^{d_c \times K}$, as shown in Equation 1,

$$z_q(v^i) = \boldsymbol{c}_k, \; k = q(v^i) = \text{argmin}_k ||v_i - \boldsymbol{c}_k||_2 \tag{1}$$

where k ($k \in [1, K]$) indexes the closest cluster (w.r.t., l_2 distance) that v_i ($i \in [1, N]$) belongs to. The corresponding output after visual discretization is $\{z_q(v^1), z_q(v^2), ..., z_q(v^N)\}$.

To learn the codebook within the framework, we consider our latent codebook space z as a random variable, and we assume a standard gaussian prior $p(z)$ over the latents. Given an image x, we also define a posterior for the latent, expressed in the bayes form,

$$p(z|x) = \frac{\int_l p(x|z,l)p(z)}{p(x)} \tag{2}$$

where l represents the language embedding space. As $p(x)$ is computationally intractable, we instead optimize a restricted family of independent gaussians $q(z|x)$. In our case, $q(z|x)$ is approximated by the visual encoder and $p(x|z,l)$ is represented by a multi-modal fusion encoder in Figure 1. By maximizing the log-likelihood of $p(x|z,l)$, we get a VAE objective. However, as the look-up operation is "discrete" and non-differentiable,

we extend a discretized VQ-VAE objective [10] and adopt gumbel-softmax [11] for gradient back-propagation.

$$\begin{aligned}\mathcal{L}_{codebook} =& \mathbb{E}(-\log(p(x|z_q(x), l) + ||\text{sg}[\text{ViT}(x)] - \boldsymbol{c}||_2^2 \\ &+ \beta||\text{ViT}(x) - \text{sg}[\boldsymbol{c}]||_2^2)\end{aligned} \tag{3}$$

where $\boldsymbol{c}$ can be considered as tokenization of visual input. Each element $\boldsymbol{c}_i$ is patchwise codeword representation for patch v^i. sg stands for stop-gradient. The first term corresponds to the pixel loss in Figure 1. The second and third term are essentially enforcing a bi-directional mapping: learning codebook vectors that align to the encoder outputs and learning encoder outputs that align to codebook vector.

B. Masked Image Modeling via Codebook

With the codebook introduced in the previous section, we build a counterpart of masked language modeling (MLM) in the BERT objective on top of "discrete" visual tokens, which we call masked image modeling (MIM). As shown in Figure 1, an input image x is first patchified into a sequence of visual patches $\{x^1, x^2, ..., x^N\}$ before fed into the visual encoder ViT. These encoded visual tokens are classified into masked V and unmasked tokens $\hat{V}$, used to calculate the MIM loss. Denote encoded text tokens as $T = \{t^1, t^2, ..., t^M\}$ (M is the length of the text),

$$\mathcal{L}_{\text{mim}} = -\mathbb{E} \log p(q(V) \,|\, l(T), z_q(\hat{V})) \tag{4}$$

where the masked token index $q(V)$ is predicted based on unmasked visual tokens and the text embeddings. The supervision signal for training $\mathcal{L}_{\text{mim}}$ comes from the codebook indices for the masked visual tokens.

C. Vision Language Pretraining

Two commonly used objectives for multimodal training frameworks are: (i) masked language modeling loss (MLM) and (ii) image-text matching (ITM), which we build on top of multi-modal fusion encoder.

Image-Text Matching (ITM) Loss Given an arbitrary pair of image and text, ITM predicts whether they are aligned (positive pairs) or not (negative pairs). This procedure can be formulated as a binary classification problem. Specifically, [CLS] token from the fusion encoder is used as the joint representation of the image-text pair. ITM head is a fully connected layer to predict the matching probability $\boldsymbol{p}_{\text{itm}}$. We assume that each image-text pair (I_i, T_i) sampled from the pre-training datasets is a positive example and construct negative examples through the following strategy: for each image I_i within the batch, we sample one negative text T_j from the same batch based on the contrastive similarity distribution. So that text that is more similar to this image will have a higher chance to get sampled. Similarly, one hard negative image will be sampled for each text T_i. We denote y_{itm} as the ground-truth labels indicating whether the image-text pair is positive or negative.

$$\mathcal{L}_{\text{itm}} = -\mathbb{E}_{I,T\sim \boldsymbol{p}_{\text{data}}} H(\boldsymbol{p}_{\text{itm}}, \boldsymbol{y}_{\text{itm}}) \tag{5}$$

where H is the cross entropy operator.

Masked Language Modeling (MLM) Loss We follow the design of MLM loss from Bert [2], which aims to predict the ground-truth labels of masked text tokens y_{mlm}. Specifically, we randomly mask out 15% of input text tokens, those masked tokens are replaced with special token [MASK]. Different from Bert, our MLM loss is conditioned on both surrounding text tokens and image representations. Assume the predicted token probability is $\boldsymbol{p}_{\text{mlm}}$, we construct the loss objective as follows,

$$\mathcal{L}_{\text{mlm}} = -\mathbb{E}_{I,\hat{T}\sim \boldsymbol{p}_{\text{data}}} H(p_{\text{mlm}}, \boldsymbol{y}_{\text{mlm}}) \tag{6}$$

where $\hat{T}$ is the text token sequence after masking.

In summary, we simultaneously optimize the codebook and the dual unimodal encoders within the framework in an end-to-end manner, employing the losses discussed in previous sections as follows,

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{codebook}} + \mathcal{L}_{\text{mim}} + \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{mlm}} \tag{7}$$

among which ITM and MLM losses are calculated with text representation as queries and visual representation as key and values. The cross-attention relation is reversed for MIM and pixel losses. The pixel loss is a part of codebook loss outlined in Equation 3. Arrows of the same color in Figure 1 indicate the flows for calculating a corresponding objective.

IV. Experiments

To evaluate our approach, we conduct extensive studies on commonly used benchmarks and present experimental comparisons against state-of-the-art V&L methods as shown in this section.

A. Pre-training Datasets

We follow previous experimental protocols [9], [18] for fair comparisons. The pretraining datasets include COCO [19], Visual Genome (VG) [20], Conceptual Captions (CC) [21], and SBU Captions [24]. In total, there are 4.0M unique images and 5.1M image-text pairs.

B. Downstream Tasks

Image-Text Retrieval This consists of two tasks: (1) image as query and text as targets (TR); (2) text as query and image as targets (IR). The pre-trained model is evaluated on COCO [19] and Flickr30K [25] by following both fine-tuning and zero-shot settings. For the fine-tuning setting, the pre-trained model is fine-tuned on the training data and evaluated on the validation/test data. For the zero-shot setting, the pre-trained model is directly evaluated on the test data without any further training.

Visual Question Answering (VQA) [26] This task aims to predict the answer given an image and a question (in text format), which requires an understanding of vision, language and commonsense knowledge to answer. We consider this task as a generation problem by following the same setting in [18].

Visual Entailment (SNLI-VE) [27] It predicts whether a given image semantically entails a given text, which is a three-classes classification problem. Specifically, the class or relationship

TABLE I
PERFORMANCE COMPARISON OF **ZERO-SHOT** IMAGE-TEXT RETRIEVAL ON FLICKR30K AND COCO DATASETS.

Method	Flickr30K (1K)						MSCOCO (5K)					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT [39]	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
Unicoder-VL [35]	64.3	85.8	92.3	48.4	76.0	85.2	-	-	-	-	-	-
UNITER [9]	80.7	95.7	98.0	66.2	88.4	92.9	-	-	-	-	-	-
ViLT [4]	73.2	93.6	96.5	55.0	82.5	89.8	56.5	82.6	89.6	40.4	70.0	81.1
CLIP [6]	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN [7]	88.6	98.7	**99.7**	75.7	93.8	**96.8**	58.6	83.0	89.7	45.6	69.8	78.6
ALBEF 4M [18]	90.5	98.8	**99.7**	76.8	93.7	96.7	68.6	89.5	94.7	50.1	76.4	84.5
CB-ViLA(Ours)	**91.9**	**99.1**	99.6	**79.1**	**94.5**	96.6	**70.1**	**90.2**	**95.3**	**52.4**	**77.8**	**86.0**

TABLE II
PERFORMANCE COMPARISON OF **FINE-TUNED** IMAGE-TEXT RETRIEVAL ON FLICKR30K AND COCO DATASETS.

Method	Flickr30K (1K)						MSCOCO (5K)					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT [39]	87.0	97.6	99.2	73.1	92.6	96.0	66.4	89.8	94.4	50.5	78.7	87.1
UNITER [9]	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
VILLA [32]	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
OSCAR [5]	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
ViLT [4]	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
UNIMO [40]	89.7	98.4	99.1	74.6	93.4	96.0	-	-	-	-	-	-
SOHO [14]	86.5	98.1	99.3	72.5	92.7	96.1	66.4	88.2	93.8	50.6	78.0	86.7
ALBEF 4M [18]	94.3	**99.4**	99.8	82.8	**96.7**	**98.4**	73.1	91.4	96.0	56.8	81.5	89.2
CB-ViLA(Ours)	**95.1**	99.1	**99.9**	**83.0**	96.1	98.2	**76.2**	**92.3**	**97.1**	**58.1**	**83.1**	**89.6**

between any given image-text pair can be entailment, neutral, or contradictory.

Visual Reasoning (NLVR^2) [28] The task determines whether a natural language caption is true about a pair of photographs. We evaluate our model on NLVR^2 dataset which contains 107,292 examples of human-written English sentences paired with web photographs.

C. Implementation Details

We use standard ViT-B/16 [15] as our vision encoder and 6-layer Bert$_{\text{base}}$ [2] as text encoder. Our multi-modal fusion encoder is another 6-layer Bert but has the same architecture as the text encoder. The multi-modal fusion encoder is shared among ITM, MLM, Pixel and MIM for forward passes and we adopt cross-attention mechanism to reduce computation. We set codebook size to 3K, as we didn't observe further improvement with larger codebook size. We follow UNITER [9] to set 15% masking ratio for masked language modeling and we follow MAE [29] to set 75% masking ratio for masked image modeling. In the pre-training stage, the model is trained for 30 epochs with a batch size of 512. We use mini-batch AdamW optimizer [30] with a weight decay of 0.02. We burn in MIM and MLM objectives after warming up the codebook for 1,000 iterations by training ITM and Pixel loss objectives. The learning rate is initialized as $1e-5$ and first warmed-up to $1e-4$ after 1,000 iterations. Then it's decreased with a cosine decay strategy to $1e-5$. All of our experiments were performed on 8 NVIDIA A100 GPUs, with an approximate training time of 60 hours.

D. Evaluation on Image-Text Retrieval

For the image-text retrieval tasks, we conduct two different scenarios for evaluation: "zero-shot" retrieval task and "after-finetuning" retrieval task, following the setting in [5], [9], [18]. We compare with both early-fusion methods such as ViLT, OSCAR, UNITER and late-fusion methods such as [7], [31]. ALBEF is a hybrid approach that performs feature alignment along with fusion. We evaluate our method on Flickr30K (1K test set) and MSCOCO (5K) in Table I and II. In the zero-shot setting, we achieve $11.2\%/12.9\%$ TR/IR improvement compared with the early-fusion approach UNITER on Flickr30K in R@1. Compared to SOTA late-fusion approach ALIGN, our approach increases $3.3\%/3.4\%$ respectively on Flickr30K R@1 and a significant $11.5\%/6.8\%$ gain on MSCOCO R@1. We hypothesize that MSCOCO is more challenging than Flickr and thus more representative of the true performance gap. Comparing to a recent SOTA approach ALBEF (4M), we show a margin of $1.4\%/2.3\%$ in terms of R@1 for TR/IR on Flickr30K and $1.5\%/2.3\%$ R@1 for TR/IR on MSCOCO respectively. In the finetuning comparison, performance tend to converge, but we observe a similar trend in performance boost across Flickr30K and MSCOCO, especially with R@1 metrics.

E. Evaluation on VQA, NLVR and VE

Following previous approaches [9], [18], we further report performances on various other vision-language tasks such as VQA, NLVR and SNLI-VE. It's worth noting that some results are not directly comparable as UNITER additionally uses out-of-

TABLE III
COMPARISON WITH A VARIETY OF STATE-OF-THE-ART METHODS ON DOWNSTREAM VISION-LANGUAGE TASKS: VQA, NVLR^2, SNLI-VE.

Method	VQA		NLVR^2		SNLI-VE	
	test-dev	test-std	dev	test-P	val	test
VisualBERT [36]	70.80	71.00	67.40	67.00	-	-
VL-BERT [17]	71.16	-	-	-	-	-
LXMERT [37]	72.42	72.54	74.90	74.50	-	-
12-in-1 [34]	73.15	-	-	78.87	-	76.95
UNITER [9]	72.70	72.91	77.18	77.85	78.59	78.28
VL-BART/T5 [38]	-	71.3	-	73.6	-	-
ViLT [4]	70.94	-	75.24	76.21	-	-
OSCAR [5]	73.16	73.44	78.07	78.36	-	-
VILLA [32]	73.59	73.67	78.39	79.30	79.47	79.03
ALBEF 4M[18]	74.54	**74.70**	80.24	80.50	80.14	80.30
CB-ViLA(Ours)	**75.84**	74.20	**80.50**	**80.84**	**81.47**	**81.40**

TABLE IV
PERFORMANCE COMPARISON OF ZERO-SHOT IMAGE-TEXT RETRIEVAL ON COCO DATASETS FOR ABLATION STUDIES.

Objective functions	MSCOCO (5K)					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
MLM+ITM	65.2	86.4	91.5	47.2	73.1	81.3
MLM+ITM+Pixel	67.4	88.2	93.9	49.5	75.9	84.4
MLM+ITM+Pixel+MIM	70.1	90.2	95.3	52.4	77.8	86.0

domain data, OSCAR leverages additional object tags and [32] with adversarial data augmentation. Nevertheless, we observe competitive performance of our method on all tasks across different datasets in Table III. Specifically, we cast VQA as an answer generation problem [18], where we finetune an auto-regressive text decoder which receives inputs from the multi-modal fusion layer. NLVR^2 and SNLI-VE are cast as binary classification and three-way classification problem respectively. The input to NLVR^2 is a pair of images and a text description, involving comprehensive pairwise relationship reasoning [28] such as co-reference, comparisons, negation, coordination etc. We find our approach performing uniformly better than prior approaches on the two tasks. Compared to Image/Text retrieval, VQA, NLVR^2 and SNLI-VE require comprehensive reasoning over the relationships between image and text descriptions, which we have explicitly modeled in Equation 3 as maximizing the reconstruction posterior conditioned on language and visual semantics. We hypothesize that MIM, together with MLM, help strengthen the model's reasoning ability between the two modalities.

F. Ablation Study

In this section, we do ablation studies on our codebook by manipulating its corresponding loss objectives. Specifically, we conduct zero-shot evaluations on the test set of MSCOCO (5K) for text and image retrieval tasks as they are more reflective of the learned representations. As shown in Table IV, the first row is our ablation study baseline which only reserves the MLM and ITM objectives, as they are commonly used in the literature. As shown in Equation 3, pixel loss contributes to codebook and visual encoder optimization. A recent approach MAE [29] shows that pixel loss is effective for improving

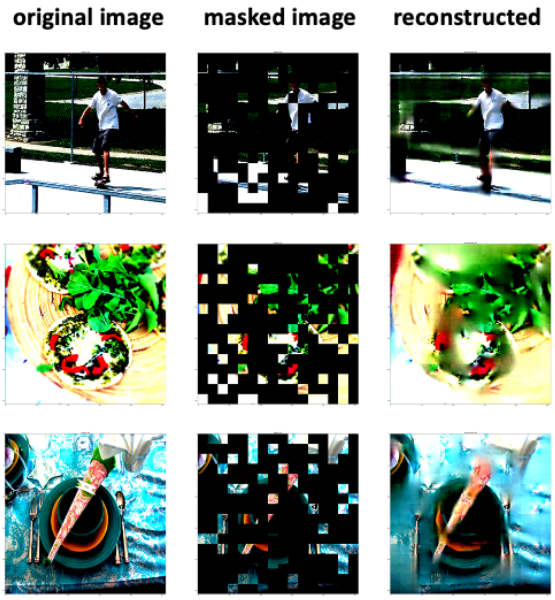

Fig. 2. Reconstruction results for the decoded masked image tokens.

unimodal encoder representation. Our experiments in row 2 also confirm the observation. The difference is that ours is a multi-modal framework and we adopt a symmetric encoding/decoding process. By adding masked image modeling on top of codebook optimization, our model accumulates an additional gain of $2.7\%/2.9\%$ in R@1 for TR and IR in the study.

G. Codebook Visualization

In this section, we answer the question: *will the learned codebook have semantic meanings?* To answer this, we randomly sample codewords in the codebook, and for each codeword, we group corresponding image patches that are discretized under this codeword index. The visualization of eight random codewords are shown in Figure 3. We observe that each codeword represents a unique pattern shared among its corresponding patches.

H. Reconstruction Results

The visualization of the reconstruction results for the masked tokens is displayed in Figure 2. We show the original image (left), the masked image (middle) and our reconstruction (right). The masking ratio is 75%, leaving only 64 out of 256 patches. The predictions differ plausibly from the original images, showing that the model can generalize. For example in the 3rd row of Figure 2, the reconstructed brush tip is pink, as opposed to green in the original image. Our hypothesis is that the model infers the color from neighboring pink patches whereas the green patches were invisible (masked) as shown in the middle.

I. Cross-attention Visualization

We visualize the cross-attention maps using Grad-CAM[33] to provide qualitative assessment of our approach. Figure 4 shows that our method is able to associate language with "regions of interest" by attending to meaningful objects and

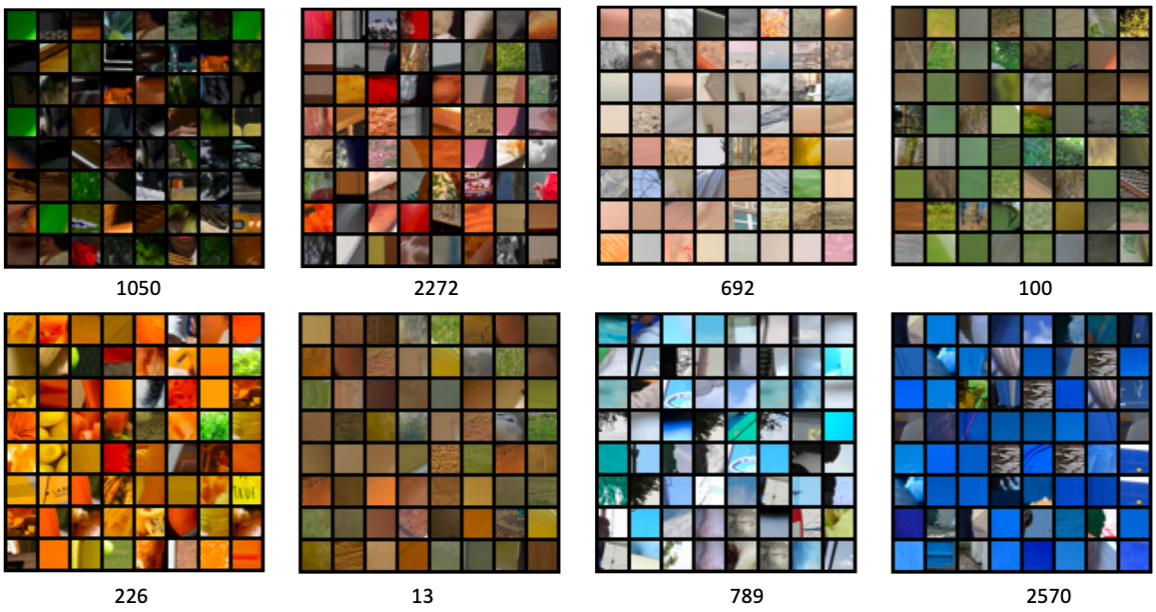

Fig. 3. Visualization of image patches that produce the same codeword in different codebooks.

Fig. 4. Grad-CAM visualization on the cross-attention maps corresponding to individual words.

locations, visually reflecting the quality of our model in multimodal alignment. For example, in Figure 4, the model is able to associate noun words such as "woman", "sunvisor", "bird", "branch" with the correct regions in the image. At the meantime, for verbs such as "swinging" and "sitting", the model is able to attend to meaningful semantics by fixating on the arms performing the action and the twig where the bird's leg touches.

V. Conclusion

Aligning signals from visual and language modalities is not easy in vision-language representation learning due to the mismatch, where visual signal is continuous and high-dimensional as opposed to language which is discretized in nature and semantics-rich. We propose **CB-ViLA** to discover the visual semantics via joint-training of a codebook, which in turn helps bridge the semantic gap by discretizing the visual signals. In this way, we symmetrically derive an equivalent of "MLM" for the vision side (i.e., MIM), which is conditioned on text and self-supervised. Ablation studies and downtream evaluations reveal that discretizing visual signals with visual semantic codebook facilitates alignment and multi-modal interactions with text signals. We hope to inspire more work in this direction.

REFERENCES

[1] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell and S. Agarwal. "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[2] J. Devlin, M.W. Chang, K. Lee and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] P. Zhang, X. Li, X. Hu, J. Yang, and L. Zhang and L. Wang and Y. Choi and J. Gao. "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[4] W. Kim, B. Son and I. Kim. "Vilt: Vision-and-language transformer without convolution or region supervision," *arXiv preprint arXiv:2102.03334*, 2021.

[5] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei and Y. Choi. "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision,* Springer 2020, August. pp. 121–137.

[6] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark and G. Krueger. "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[7] C. Jia, Y. Yang, Y. Xia, Y.T. Chen, Z. Parekh, H. Pham, Q.V. Le, Y. Sung, Z. Li and T. Duerig. "Scaling up visual and vision-language representation learning with noisy text supervision," *arXiv preprint arXiv:2102.05918*, 2021.

[8] A.V.D. Oord, Y. Li and O. Vinyals. "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[9] Y.C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng and J. Liu. "Uniter: Universal image-text representation learning", in *European conference on computer vision*, Springer 2020 August, pp. 104–120.

[10] A.V.D. Oord, O. Vinyals and K. Kavukcuoglu. "Neural discrete representation learning", *arXiv preprint arXiv:1711.00937*, 2017.

[11] E. Jang, S. Gu and B. Poole. "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[12] H.B. Bao, L. Dong, and F.R. Wei. "BEiT: BERT Pre-Training of Image Transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever. "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.

[14] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu. "Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12976–12985.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and J. Uszkoreit. "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei and J. Dai. "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[17] J. Lu, D. Batra, D. Parikh and S. Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *arXiv preprint arXiv:1908.02265*, 2019.

[18] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong and S.C.H. Hoi. "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in Neural Information Processing Systems*, 2021.

[19] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C.L. Zitnick. "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer 2014, pp. 740–755.

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.J. Li, D.A. Shamma and M.S. Bernstein. "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, Springer 2017, vol. 123, num. 1, pp. 32–73.

[21] P. Sharma, N. Ding, S. Goodman and R. Soricut. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[22] Yang, Jinyu and Duan, Jiali and Tran, Son and Xu, Yi and Chanda, Sampath and Chen, Liqun and Zeng, Belinda and Chilimbi, Trishul and Huang, Junzhou. "Vision-Language Pre-Training with Triple Contrastive Learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[23] Duan, Jiali and Lin, Yen-Liang and Tran, Son and Davis, Larry S and Kuo, C-C Jay. "Slade: A self-training framework for distance metric learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[24] V. Ordonez, G. Kulkarni and T. Berg. "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, 2011, pp. 1143–1151.

[25] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier, and S. Lazebnik. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015, pp. 2641–2649.

[26] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh. "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[27] N. Xie, F. Lai, D. Doran and A. Kadav. "Visual entailment: A novel task for fine-grained image understanding," *arXiv preprint arXiv:1901.06706*, 2019.

[28] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai and Y. Artzi. "A corpus for reasoning about natural language grounded in photographs," *arXiv preprint arXiv:1811.00491*, 2018.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick. "Masked Autoencoders Are Scalable Vision Learners," *arXiv preprint arXiv:2111.06377*, 2021.

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

[31] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever. "Improving language understanding by generative pre-training," 2018.

[32] Z. Gan, Y.C. Chen, L. Li, C. Zhu, Y. Cheng and J. Liu. "Large-scale adversarial training for vision-and-language representation learning," *arXiv preprint arXiv:2006.06195*, 2020.

[33] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision (CVPR)*, 2017, pp. 618–626.

[34] J. Lu, V. Goswami, M. Rohrbach, D. Parikh and S. Lee. "12-in-1: Multi-task vision and language representation learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10437–10446.

[35] G. Li, N. Duan, Y. Fang, M. Gong and D. Jiang. "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, num. 07, pp.11336–11344.

[36] L.H. Li, M. Yatskar, D. Yin, C.J. Hsieh and K.W. Chang. "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[37] Hao Tan and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[38] J. Cho, J. Lei, H. Tan and M. Bansal. "Unifying vision-and-language tasks via text generation," *arXiv preprint arXiv:2102.02779*, 2021.

[39] D. Qi, L. Su, J. Song, E. Cui, T. Bharti and A. Sacheti. "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," *arXiv preprint arXiv:2001.07966*, 2020.

[40] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang. "Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning," *arXiv preprint arXiv:2012.15409*, 2020.

[41] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. "Videobert: A joint model for video and language representation learning." *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2019, pp. 7464-7473.

[42] T. Fu, L. Li, Z. Gan, K. Lin, W.Y. Wang, L. Wang, and Z. Liu. "VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling." *arXiv preprint arXiv:2111.12681*, 2021.

[43] X. Dong, et al. "PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers." *arXiv preprint arXiv:2111.12710*, 2021.

[44] Duan, Jiali and Chen, Liqun and Tran, Son and Yang, Jinyu and Xu, Yi and Zeng, Belinda and Chilimbi, Trishul. "Multi-modal Alignment using Representation Codebook." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.