arXiv:2210.05335v3 [cs.CV] 20 Jul 2023

MAP: Multimodal Uncertainty-Aware Vision-Language Pre-training Model

Yatai Ji[1*] Junjie Wang[2*] Yuan Gong[1] Lin Zhang[3] Yanru Zhu[1]
Hongfa Wang[4] Jiaxing Zhang[3] Tetsuya Sakai[2†] Yujiu Yang[1†]

[1]Tsinghua University [2]Waseda University [3]IDEA [4]Tencent TEG

{jyt21, gong-y21, zhuyr20}@mails.tsinghua.edu.cn yang.yujiu@sz.tsinghua.edu.cn
wjj1020181822@toki.waseda.jp tetsuyasakai@acm.org
{zhanglin, zhangjiaxing}@idea.edu.cn hongfawang@tencent.com

Abstract

Multimodal semantic understanding often has to deal with uncertainty, which means the obtained messages tend to refer to multiple targets. Such uncertainty is problematic for our interpretation, including inter- and intra-modal uncertainty. Little effort has studied the modeling of this uncertainty, particularly in pre-training on unlabeled datasets and fine-tuning in task-specific downstream datasets. In this paper, we project the representations of all modalities as probabilistic distributions via a Probability Distribution Encoder (PDE) by utilizing sequence-level interactions. Compared to the existing deterministic methods, such uncertainty modeling can convey richer multimodal semantic information and more complex relationships. Furthermore, we integrate uncertainty modeling with popular pre-training frameworks and propose suitable pre-training tasks: Distribution-based Vision-Language Contrastive learning (D-VLC), Distribution-based Masked Language Modeling (D-MLM), and Distribution-based Image-Text Matching (D-ITM). The fine-tuned models are applied to challenging downstream tasks, including image-text retrieval, visual question answering, visual reasoning, and visual entailment, and achieve state-of-the-art results.

1. Introduction

Precise understanding is a fundamental ability of human intelligence, whether it involves localizing objects from similar semantics or finding corresponding across multiple modalities. Our artificial models suppose to do the same, pinpointing exact concepts from rich multimodal semantic scenarios. However, this kind of precise understanding is challenging. Information from different modalities can present rich semantics from each other, but the resulting ambiguity and noise are also greater than the case with a single modality.

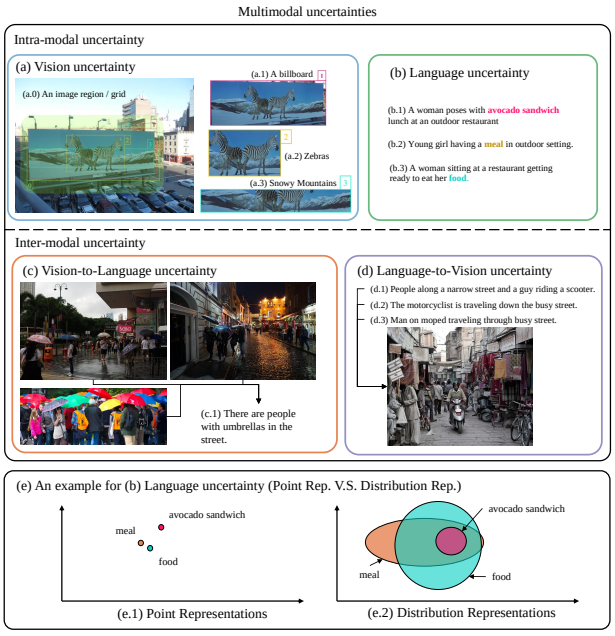


Figure 1. Multimodal uncertainties and an example for language uncertainty (b) by modeling as point representations and distribution representations. The images and text are from MSCOCO [30].

Multimodal representation learning methods hold the promise of promoting the desired precise understanding across different modalities [13]. While these methods have shown promising results, current methods face the challenge of uncertainty [7, 51], including within a modality and between modalities. Considering image (a.0) in Fig. 1 as an example, one vision region includes multiple objects, such as a billboard, several zebras and others. Therefore, it is unclear which objects when mentioning this region. In the language domain, the complex relationships of words lead to uncertainty, such as synonymy and hyponymy. In Fig. 1 (c)&(d), the same object often has different descriptions

*Equal contribution.
†Corresponding Author.

from different modalities, such as text and images, which manifests inter-modal uncertainty. Instead, previous methods often neglect the uncertainty [11, 19, 46], resulting in limited understanding ability on complicated concept hierarchies and poor prediction diversity. Therefore, it is desirable to model such uncertainty.

Moreover, with multimodal datasets becoming more commonplace, there is a flourishing trend to implement pre-training models, particularly Vision-Language Pre-training (VLP), to support downstream applications [6, 18, 23, 36, 50]. Existing deterministic representations, however, often fail to understand uncertainty in pre-training data, as they can only express positions in semantic space and measure the relationship between targets in certainty, such as Euclidean distance. *How can we efficiently model uncertainty in multi-modalities when dealing with pre-training models?*

Applying Gaussian distribution is one of the prominent approaches used for modeling uncertainty in the representation space [40, 45, 51, 54]. In these methods, however, the obtained uncertainty depends on individual features rather than considering the whole features together, which ignores the inner connection between features. To exploit this connection, we implicitly model them when formulating the uncertainty with a module called Probability Distribution Encoder (PDE). Inspired by the self-attention mechanism [44], we further add the interaction between text tokens and image patches when constructing our distribution representations to capture more information. In Figure 1 (e), we provide an example for two different types of representations to describe the language uncertainty, where the distribution representations can express richer semantic relationships than the conventional point representations. The distribution variance measures the uncertainty of the corresponding text. As a byproduct, distribution representations enable diverse generations, providing multiple reasonable predictions with random sampling.

In this paper, we integrate this uncertainty modeling in the pre-training framework, resulting in three new tasks: Distribution-based Vision-Language Contrastive learning (D-VLC), Distribution-based Masked Language Modeling (D-MLM), and Distribution-based Image-Text Matching (D-ITM) pre-training tasks. All these tasks are to deal with cross-modality alignment. More specifically, D-VLC is to handle the coarse-grained cross-modal alignment, which measures the whole distributions to align representations from different domains. D-MLM and D-ITM are implemented after the fine-grained interaction between different modalities, providing the token level and overall level alignment for images and text.

Our contributions are summarized as follows:
1) We focus on the semantic uncertainty of multimodal understanding and propose a new module, called Probability Distribution Encoder, to frame the uncertainty in multimodal representations as Gaussian distributions.
2) We develop three uncertainty-aware pre-training tasks to deal with large-scale unlabeled datasets, including D-VLC, D-MLM, and D-ITM tasks. To the best of our knowledge, these are the first attempt to harness the probability distribution of representations in VLP.
3) We wrap the proposed pre-training tasks into an end-2-end **M**ultimodal uncertainty-**A**ware vision-language **P**re-training model, called MAP, for downstream tasks. Experiments show MAP gains State-of-The-Art (SoTA) performance. Our code is available at https://github.com/IIGROUP/MAP.

2. Related Works

2.1. Probability Distribution Representations

Current popular representation learning methods extract features as point representations and focus on searching for the closest position to ground truth in high-level representation space. However, there is usually more than one suitable point representation, which shows the uncertainty in multiple tasks. To address this problem, the following researchers introduced probability distribution representations to infer diversely and improve robustness, avoiding model overfitting to one single solution. In the Natural Language Processing (NLP) field, multivariate Gaussian distribution was utilized to represent words [45] due to the powerful capability for representing the asymmetric relations among words. Since then, different distribution families were exploited for word representations [2, 28]. In Computer Vision (CV), for modeling vision uncertainty, some researchers introduce Gaussian representations into specific tasks, such as face recognition [4], person re-identification [54], 3D skeleton action representation [40] and pose estimation [42]. For solving the long-tail problem in relation prediction, Gaussian distribution was utilized to build objects relationship in scene graph generation [52]. Recently, constructing distributions achieved some progress to yield diverse predictions for cross-modal retrieval in multimodal field [7]. However, those existing methods only consider the feature level to build the distributions for a whole image or sentence. In this work, we model not only the whole image or sentence to the distribution representations but also each token of them, such as patches and words. Furthermore, our approach learns the multimodal uncertainty from sequence-level and feature-level interactions.

2.2. Vision-Language Pre-training (VLP)

Inspired by the Transformer structure [44] and pre-training tasks from BERT [8], the recent emergence of vision-language pre-training tasks and models have been explored to learn multimodal representations. The main process is first to pre-train the models by exploiting auxiliary tasks to understand hidden supervision information from large-scale unlabeled data. Then, the pre-trained models embed real-world objects into multimodal representations.

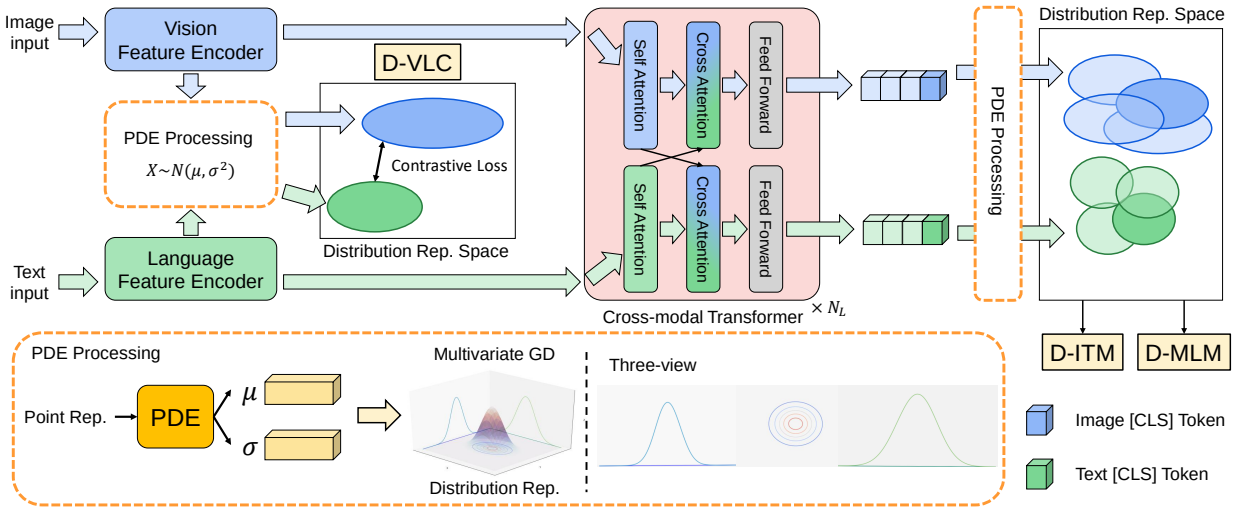


Figure 2. Pre-training model architecture and objectives of MAP. We propose PDE to model distribution representations as multivariate Gaussian Distributions (GD). "N_L" indicates the layer number of the cross-modal transformer. We perform two-dimensional Gaussian distribution as an example.

With effective universal multimodal representations, they can achieve good performance by fine-tuning on relatively small labeled datasets of VL downstream tasks. The key challenge of VLP is to design suitable pre-training objectives. Recently, mainstream strategies include Masked Language Modeling (MLM) [15, 16, 20, 23, 26], Image-Text Matching (ITM) [15, 16, 23, 26] and Vision-Language Contrastive learning (VLC) [18, 20, 26, 36]. MLM in VLP requires the model to predict the masked language context tokens by the rest of the language context tokens and vision context tokens. To understand alignment information of language context and vision context, ITM requires the model to judge whether the input of different modalities matches or not. VLC learns the similarity from inter-modal information and aligns point representations of different modalities. However, those methods only are designed in the point representation space without considering multimodal uncertainty. Therefore, we propose the D-VLC, D-MLM and D-ITM to pre-train our model in the distribution representation space. The details will be explained in Sec. 3.2.

3. Approaches

In this section, we introduce our proposed PDE and the architecture of MAP (Sec. 3.1), and the overall structure is described in Fig. 2. The details of our proposed distribution-based pre-training tasks are presented in Sec. 3.2. In addition, we further discuss the ethical considerations in Appendix C.

3.1. Model Overview

3.1.1 Probability Distribution Encoder (PDE).

The input features of PDE are from the point representation space of different modalities. To model the multimodal uncertainty, we further frame the input features as multivariate Gaussian distributions. Specifically, PDE predicts a mean vector (μ) and a variance vector (σ^2) for each input feature. The mean vector represents the center position of distributions in probabilistic space, and the variance vector expresses the scope of distributions in each dimension.

As shown in Fig. 3, we propose a probability distribution encoder (PDE) while considering that modeling the mean and variance vectors takes feature-level and sequence-level interactions. Specifically, Feed Forward layer is used for feature-level interactions and Multi-Head (MH) operation is responsible for sequence-level interactions. By applying the MH operation, the input hidden states $H \in \mathbb{R}^{T \times D}$ are split into k heads, where T is sequence length and D is hidden size. In each head, we split the features and send them to two paths (μ, σ^2). In each path, the input hidden states $H^{(i)} \in \mathbb{R}^{T \times D/2k}$ are projected to $Q^{(i)}$, $K^{(i)}$, $V^{(i)}$ in i-th head. As an example, the operation in the μ path is:

$$
\begin{aligned}
[Q_\mu^{(i)}, K_\mu^{(i)}, V_\mu^{(i)}] &= H_\mu^{(i)} W_{qkv}\,, \\
Head_\mu^{(i)} &= \mathrm{Act}\left(Q_\mu^{(i)} {K_\mu^{(i)}}^\top / \sqrt{d_k}\right) V_\mu^{(i)}\,, \\
\mathrm{MH}_\mu &= \mathrm{concat}_{i \in [k]} \left[Head_\mu^{(i)}\right] W_O\,,
\end{aligned}
\tag{1}
$$

where d_k is set to $D/(2k)$. The weight $W_{qkv} \in \mathbb{R}^{d_k \times 3d_k}$ is to project the inputs in the subspace of each head. The weight $W_O \in \mathbb{R}^{kd_k \times D}$ projects the concatenation of k head results to the output space. The "Act" includes an activation function and a normalization function for considering sequence-level interaction. The σ^2 path is similar to the μ path. Since the input point representation correlates with the mean vector, an add operation is employed to learn the mean vector. The motivations of design choices are in Sec. 4.3.2. After PDE, each vision or language token is represented as a

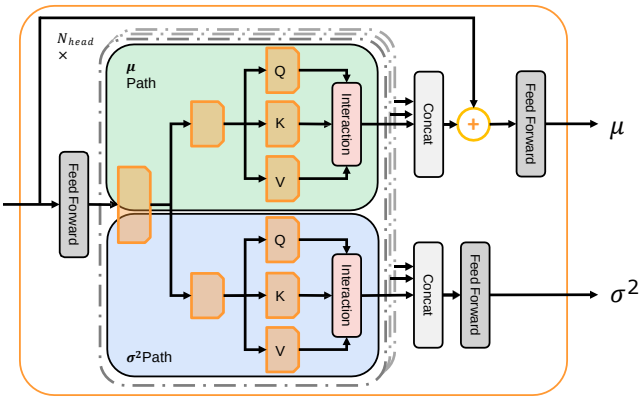


Figure 3. The architecture of Probability Distribution Encoder (PDE) block.

gaussian distribution in high-dimension probabilistic space.

3.1.2 Feature Extraction.

To extract features, we utilize an image encoder and a language encoder. In detail, we employ CLIP-ViT [36] as the image encoder and RoBERTa-Base [31] as the language encoder. An image is encoded as an patch feature sequence $\{v_{[\texttt{CLS}]}, v_1, \ldots, v_N\}$, where $v_{[\texttt{CLS}]}$ is the overall feature. Moreover, the input text is embedded into a sequence of tokens $\{w_{[\texttt{CLS}]}, w_1, \ldots, w_M\}$.

3.1.3 Cross-modal Transformer.

Recently, there are two main types of the multimodel transformer to fuse the different modalities: single-stream [6, 39, 56] and dual-stream [26, 32, 43] models.

In our method, the image patch sequences are much longer than text sequences, making the weights of vision features too large to compute the attention scores together. To address this issue, we choose the dual-stream module with two transformer branches, where self-attention scores are calculated separately.

As shown in Fig. 2, the main structure has N_L layers of cross-modal encoders. Each encoder mainly consists of two Self-Attention (SA) blocks and two Cross-Attention (CA) blocks. In the SA block of each modality, query, key and value vectors are all linearly projected from vision or language features. In the vision-to-language cross-attention block of i-th layer, query vectors represent language feature T_i' after the self-attention block, and key/value vectors denote vision feature I_i'. By employing the Multi-Head Attention (MHA) operation, the CA block enables language features to learn visual information across modalities. The language-to-vision CA block is similar to the vision-to-language one. The workflow of i-th layer encoder with SA and CA is as follows:

$$\begin{aligned} SA_{\text{vision}} &: I_i' = \text{MHA}(I_{i-1}, I_{i-1}, I_{i-1}), \\ SA_{\text{language}} &: T_i' = \text{MHA}(T_{i-1}, T_{i-1}, T_{i-1}), \\ CA_{\text{vision}} &: I_i = \text{MHA}(I_i', T_i', T_i'), \\ CA_{\text{language}} &: T_i = \text{MHA}(T_i', I_i', I_i'). \end{aligned} \tag{2}$$

For the overall structure design of MAP, we apply PDEs after feature extractors and cross-modal transformer, respectively. PDE after the feature extractor learns unimodal distribution representations to conduct the D-VLC pre-training task. PDE at the end of MAP is responsible for D-MLM, D-ITM and downstream tasks.

3.2. Distribution-based Pre-Training Tasks

In order to learn the multimodal uncertainty in common sense, we pre-train our model with distribution-based pre-training tasks on large-scale datasets.

3.2.1 Coarse-grained Pre-training.

We propose Distribution-based Vision-Language Contrastive Learning, called **D-VLC**, to realize coarse-grained semantic alignment of overall unimodal distributional representations before fusion. We compute the 2-Wasserstein distance [21, 22, 33] to measure the distance between multivariate Gaussian distributions. For two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, their 2-Wasserstein distance is defined as:

$$D_{2W} = ||\mu_1 - \mu_2||_2^2 + \text{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}). \tag{3}$$

In our modeled distributions, Σ_1 and Σ_2 are both diagonal matrices, which indicates $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} = \Sigma_1\Sigma_2$. The above formula can be rewritten as:

$$\begin{aligned} D_{2W} &= ||\mu_1 - \mu_2||_2^2 + \text{Tr}((\Sigma_1^{1/2} - \Sigma_2^{1/2})^2) \\ &= ||\mu_1 - \mu_2||_2^2 + ||\sigma_1 - \sigma_2||_2^2 , \end{aligned} \tag{4}$$

where σ refers to a standard deviation vector. The overall unimodal features denote the distribution representations of [CLS] from the PDEs following single-modal feature extractors. The similarity between an image and text is given by:

$$s(I, T) = a \cdot D_{2W}(v_{[\texttt{CLS}]}, w_{[\texttt{CLS}]}) + b , \tag{5}$$

where a is a negative scale factor since similarity is inversely proportional to the distance, and b is a shift value. For N image-text pairs in a batch, there are N positive matched samples and $N(N-1)$ negative samples. We use InfoNCE loss as follows:

$$\begin{aligned} \mathcal{L}_{NCE}^{I2T}(i) &= -\log \frac{\exp(s(I_i, T_i)/\tau)}{\sum_{n=1}^{N} \exp(s(I_i, T_n)/\tau)} , \\ \mathcal{L}_{NCE}^{T2I}(i) &= -\log \frac{\exp(s(T_i, I_i)/\tau)}{\sum_{n=1}^{N} \exp(s(T_i, I_n)/\tau)} , \end{aligned} \tag{6}$$

where τ is a learned temperature parameter. The above are summed as D-VLC loss $\mathcal{L}_{D\text{-}VLC}$.

3.2.2 Fine-grained Pre-training.

After the cross-modal transformer with fine-grained interaction on each token of different modalities, our proposed Distribution-based Masked Language Modeling (D-MLM) and Distribution-based Image Text Matching (D-ITM) can assist the model in learning fine-grained cross-modal alignment.

D-MLM requires the model to predict the masked words by understanding the text with an image. Specifically, the input text tokens are replaced by [MASK] with a probability of 15% (Details in Appendix A.3). For conducting classification in the word list to predict the original words, we sample the point vectors from distribution representations. D-MLM minimizes a Cross-Entropy (CE) loss for μ point and other sample point vectors:

$$\mathcal{L}_{D\text{-}MLM} = \frac{1}{K+1}(\mathrm{CE}(\phi(\mu), y) + \sum_{i=1}^{K} \mathrm{CE}(\phi(z^{(i)}), y)), \quad (7)$$

where K is the sample number from gaussian distributions and y serves as a masked word label. μ is a mean vector and $z^{(i)}$ refers to stochastic sample point vectors; then, they are fed to MLM classifier ϕ. During the inference process, the final output is the mean pooling of all samples' prediction results:

$$P = \frac{1}{K+1}(\phi(\mu) + \sum_{i=1}^{K} \phi(z^{(i)})). \quad (8)$$

D-ITM provides a binary classification that predicts whether a pair of image-text is matched or not. Specifically, we sample the point vectors from $w_{\texttt{CLS}}$ distributions of vision and language features and concatenate them as the fusion features to generate the prediction.

$$\begin{aligned}\mathcal{L}_{D\text{-}ITM} = \frac{1}{K+1}(&\mathrm{CE}(\phi(\mathrm{concat}[v_\mu, w_\mu]), y) \\ &+ \sum_{i=1}^{K} \mathrm{CE}(\phi(\mathrm{concat}[v^{(i)}, w^{(i)}]), y)),\end{aligned} \quad (9)$$

where v_μ, t_μ are mean vectors of vision and language [CLS] distributions. $v^{(i)}$, $w^{(i)}$ are sample points and ϕ is the D-ITM classifier. The image-text pairs in the datasets serve as positive examples, and negative examples are constructed by randomly replacing images or text descriptions.

However, random sampling increases training difficulty. When the model is trained only with the aforementioned losses, it will lead to variance collapse. Since all sample point vectors will converge to the optimal position, the distribution representations eventually degenerate into point representations, resulting in losing the ability to learn multimodal uncertainty. Therefore, we append a regularization loss to prevent the uncertainty level of distributions is lower than a certain threshold:

$$\mathcal{L}_{reg} = \max(0, \gamma - h(\mathcal{N}(\mu, \sigma^2))), \quad (10)$$

where γ is a set threshold, which affects the uncertainty level of distributions. $h(\mathcal{N}(\mu, \sigma^2))$ is the entropy of a multivariate Gaussian distribution, which should be defined as:

$$h(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\log(\det(2\pi e\Sigma)), \quad (11)$$

where Σ is the covariance matrix, which is a diagonal matrix in our method. Therefore, the diagonal vector of Σ is σ^2 and the Eq. (11) can be transformed to:

$$\begin{aligned}h(\mathcal{N}(\mu, \sigma^2)) &= \frac{1}{2}\sum_{i=1}^{d}\log(2\pi e \cdot \sigma_i^2) \\ &= \frac{d}{2}(\log(2\pi)+1) + \sum_{i=1}^{d}\log\sigma_i\,,\end{aligned} \quad (12)$$

where d is the feature dimension.

Note that the sampling operation for $\mathcal{N}(\mu, \sigma^2)$ causes the problem of preventing gradients from propagating back. By applying the *reparameterization trick* [24], we first sample a random noise ϵ from standard normal distributions, instead of directly sampling from $\mathcal{N}(\mu, \sigma^2)$:

$$z = \mu + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \quad (13)$$

After Eq. (13), the output z obeys the predicted distributions from the PDE. Therefore, we can separate the calculations of the mean and standard deviation from the sampling operation and they are trainable.

3.2.3 Training Objectives.

During pre-training phase, the model will propagate forward three times at one step with conducting D-MLM, D-ITM and D-VLC tasks separately. Therefore, the full pre-training objective is given by:

$$\mathcal{L}_{pre} = \mathcal{L}_{D\text{-}MLM} + \mathcal{L}_{D\text{-}ITM} + \mathcal{L}_{D\text{-}VLC} + \alpha\mathcal{L}_{reg}\,, \quad (14)$$

where α is its weight.

4. Experiments

4.1. Experimental settings

By following a popular setting [9], we set all hidden feature sizes as 768, and the head number as 12 in MHA. Unless otherwise specified, the layer number (N_L) of the cross-modal transformer is set to 6. As for data processing, we resize and crop each image into the size of 384×384. The size P of the image patch is 16. And the maximum length of input text dealt is set to 50. In PDE, the head number k is set to 6 and the default "Act" function in Eq. (1) is Softmax.

For pre-training, we pre-train our model with D-MLM, D-ITM and D-VLC. The pre-training datasets include MSCOCO [30], Visual Genome (VG) [25], SBU [34] and Conceptual Captions (CC-3M) [38]. Specifically, we resize

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Group 2: Pre-training datasets include > 10M images												
ALBEF (14M) [26]	60.7	84.3	90.5	77.6	94.3	97.2	85.6	97.5	98.9	95.9	99.8	100.0
Group 1: Pre-training datasets include < 10M images												
UNITER-Large [6]	52.9	79.9	88.0	65.7	88.6	93.8	75.6	94.1	96.8	87.3	98.0	99.2
VILLA-Large [10]	-	-	-	-	-	-	76.3	94.2	96.8	87.9	97.5	98.8
UNIMO-Large [27]	-	-	-	-	-	-	78.0	94.2	97.1	89.4	98.9	99.8
VinVL-Large [56]	58.8	83.5	90.3	75.4	92.9	96.2	-	-	-	-	-	-
ViLT [23]	42.7	72.9	83.1	61.5	86.3	92.7	64.4	88.7	93.8	83.5	96.7	98.6
UNITER-Base [6]	50.3	78.5	87.2	64.4	87.4	93.1	72.5	92.4	96.8	85.9	97.1	98.8
ALBEF (4M) [26]	56.8	81.5	89.2	73.1	91.4	96.0	82.8	96.7	98.4	94.3	99.4	99.8
TCL [53]	59.0	83.2	89.9	75.6	92.8	96.7	**84.0**	96.7	98.5	**94.9**	99.5	99.8
METER [9]	57.1	82.7	90.1	76.2	93.2	96.8	82.2	96.3	98.4	94.3	**99.6**	**99.9**
MAP (ours)	**60.9**	**86.2**	**93.1**	**79.3**	**94.8**	**97.6**	83.8	**97.2**	**98.7**	**94.9**	99.5	99.8

Table 1. An overall comparison with SoTA models on fine-tuned image-text retrieval tasks. The best scores are in **bold** and the second best scores are underlined.

and crop each image into the size of 288×288. Please find more pre-training and fine-tuning details in Appendix B.1.

In all experiments, we employ the randomized Tukey HSD p-values and effect sizes based on one-way ANOVA [37] to support the statistical significance of all results (Please refer to Appendix B.7 for more details).

4.2. Results of VL Downstream Tasks

In this section, we apply our pre-trained MAP on the following 4 VL downstream tasks with 5 widely-used datasets. Image retrieval task (MSCOCO [30] and Flickr30K [35]) aims to understand the multimodal uncertainty which results from the multiplicity of concepts in images and text. This is similar to the objective of our uncertainty modeling in nature. Meanwhile, visual question answering (VQA2.0 [12]), visual reasoning (NLVR2 [41]) and visual entailment (SNLI-VE [49]) implicitly perform ambiguous semantics in unimodal and cross-modal items. Therefore, we further evaluate our MAP on the aforementioned tasks to varying the effectiveness and generalization ability of uncertainty modeling. For fair experimental environments, we group previous models with different sizes of pre-training datasets. Please find more details of the datasets, model descriptions and additional experiments in Appendix B.

4.2.1 Evaluation on Image-Text Retrieval

As shown in Table 1, our MAP achieves the best performance on MSCOCO and gains either the best or the second-best scores on Flickr30K. Specially, while ALBEF has specially-designed objectives for retrieval, the MAP also outperforms ALBEF (14M pre-training images) in all metrics on the MSCOCO retrieval task. The results show the effectiveness and advantages of uncertainty modeling. For the Flickr30K dataset, our MAP achieves the best performance or only about 0.1 point behind the best score. PCME also utilizes probabilistic distribution representations to conduct retrieval task, and we show the comparison in Appendix B.6.

Model	VQA2.0	NLVR2		SNLI-VE	
	test-dev	dev	test-p	val	test
Group 2: Pre-training datasets include >10M images (Base size)					
ALBEF (14M) [26]	75.84	81.72	81.77	84.20	84.15
SimVLM-Base [48]	77.87	82.55	83.14	80.80	80.91
Group 1: Pre-training datasets include <10M images (Base size)					
ViLT [23]	71.26	75.70	76.13	-	-
UNITER-Base [6]	72.70	77.18	77.85	78.59	78.28
OSCAR-Base [29]	73.16	78.07	78.36	-	-
UNIMO-Base [27]	73.79	-	-	80.00	79.10
ALBEF (4M) [26]	74.54	80.24	80.50	80.14	80.30
VinVL-Base [56]	75.95	82.05	83.08	-	-
VLMo-Base [47]	76.64	82.77	83.34	-	-
METER [9]	77.68	82.33	83.05	80.86	81.19
MAP (ours)	**78.03**	**83.30**	**83.48**	**81.40**	**81.39**

Table 2. An overall comparison with SoTA models on visual question answering, visual reasoning, visual entailment tasks. The best scores are in **bold** and the second best scores are underlined.

4.2.2 Evaluation on VQA2.0, NVLR2, and SNLI-VE

As shown in Table 2, our MAP outperforms the previous SoTA models in Group 1. Compared to VLMo-Base, the MAP improves 0.53 points on NLVR2 dev. Our model brings the +0.35 points improvement on VQA2.0 test-dev and +0.54 points performance gains on SNLI-VE val over METER. Notably, MAP outperforms SimVLM-Base (1.8B pre-training images) in all tasks, which further demonstrates

	VQA2.0	SNLI-VE		NLVR2	
	test-dev	val	test	dev	test-p
Random initialization					
MAP w/o PDE	72.09	75.91	76.28	50.86	51.07
MAP	73.35	76.67	76.86	51.12	51.07
Pretained on MSCOCO					
MAP w/o PDE	74.57	79.42	79.84	77.72	79.31
MAP	**75.01**	**80.05**	**80.31**	**78.96**	**79.64**

Table 3. The effectiveness of probability distribution representations on VL downstream tasks. For "MAP w/o PDE", we train a new model without PDE to conduct the experiments. Pre-trained methods for MAP: D-MLM, D-ITM. Pre-trained methods for MAP w/o PDE: MLM, ITM.

the effectiveness of uncertainty modeling.

4.3. Ablation Studies

4.3.1 How do the probability distribution representations affect VL downstream task?

As shown in Table 3, applying PDE helps the model to achieve a better performance, which matters significantly in VL downstream tasks. In both random and pre-trained weights initialization cases, distribution representations gain a better capability of VL understanding than the point representations, which is because the distribution representations can express richer semantics by learning multimodal uncertainty.

4.3.2 How does the structure of PDE behave?

We remove the sequence-level interaction in PDE and call it "MLP only" (MultiLayer Perceptron), which is the common method in the previous works [7, 51, 54]. Table 4 shows that PDE (Softmax) outperforms "MLP only" in VQA2.0, which benefits from the sequence-level information. Moreover, we design several candidate activation functions: ReLU, ReLU^2, Sigmoid, and Softmax. Notably, "MLP only" outperforms "ReLU" and "ReLU^2", which demonstrates that it is important to consider how to design the sequence-level interaction. The function Sigmoid projects the input values between 0 to 1, which smoothly assigns weights between different tokens. The function Softmax outperforms the others in VQA2.0, which implies that Softmax is suitable to express the correlation between tokens. Therefore, we set Softmax as the default activation function in sequence-level interaction.

4.3.3 What is the performance of different pre-training objectives?

Table 5 presents that different choices of pre-training tasks affect the VL downstream tasks performance. According to the chart, results without D-MLM pre-training are the worst in all pre-training strategies, which means D-MLM plays the most important role in pre-training. Both D-VLC and D-ITM assist the model in learning semantic similarity between vision and language modality. In VQA2.0, D-VLC makes a larger improvement than D-ITM, whereas D-ITM is more effective than D-VLC in SNLI-VE and NLVR2.

Structure		VQA2.0 (test-dev)
MLP only		72.01
PDE	ReLU+Normal	69.70
	ReLU^2+Normal	70.53
	Sigmoid+Normal	73.34
	Softmax	**73.35**

Table 4. Effect of different structures of PDE. We explore the different designs of "Act" in Equation 1. Normal denotes the normalization operation.

Training strategies	VQA2.0 test-dev	SNLI-VE test-p	NLVR2 test
Random Initialization	73.35	76.86	51.07
D-MLM, D-ITM	75.01	80.31	**79.64**
D-MLM, D-VLC	75.06	80.12	77.90
D-ITM, D-VLC	71.02	78.54	73.64
D-MLM, D-ITM, D-VLC	**75.16**	**80.39**	79.47

Table 5. The effect of distribution-based pre-training tasks. We pre-train the model on the MSCOCO dataset.

Layer Number	Random Initializing	Pre-training
2	72.71	73.78
4	73.32	74.73
6	**73.35**	75.16
8	73.31	**75.26**

Table 6. The effect of different layer numbers in the cross-modal transformer on VQA2.0.

4.3.4 Does the number of layers of cross-modal transformer matter?

As shown in Table 6, we explore the effect of layer number on VQA2.0 by random initializing or pre-training with D-MLM, D-ITM, D-VLC on the MSCOCO dataset. By random initializing, the model with six layers achieves the best performance, which encounters a bottleneck. After pre-training, the model with eight layers makes little progress from six layers, which implies that pre-training helps the model break the aforementioned bottleneck of parameters. The reason perhaps is that pre-training with large-scale data alleviates the problem of over-fitting by more parameters. Moreover, as the layer number decreases, the effect of pre-training will

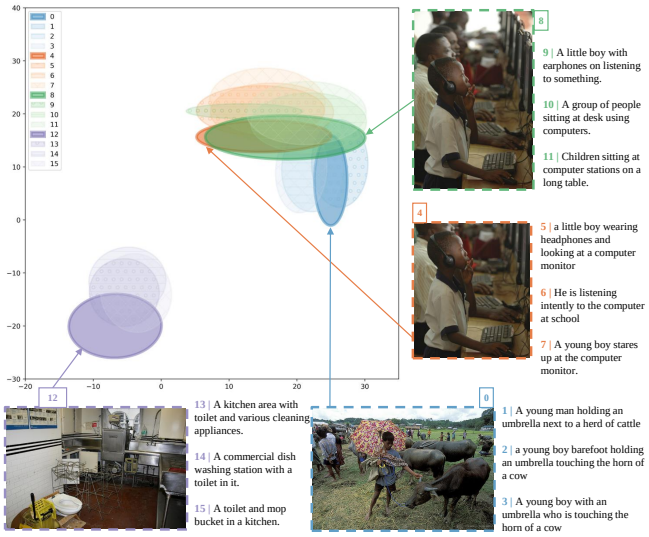


Figure 4. Visualization of the distribution representations from pre-trained MAP[2]. Each 2D Gaussian distribution is represented as an ellipse with 95% confidence. The labels of images and related captions are in the legend.

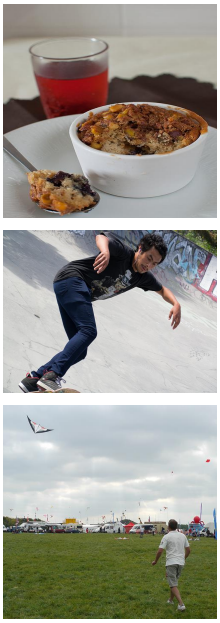
reduce due to the model's limited learning capacity.

4.4. Uncertainty Modeling Analysis

Visualization. To perform visualization analysis for distribution representations from pre-trained MAP, we conduct 2D toy experiments for the images and related descriptions. Fig. 4 shows the behaviors of the distribution representations, which present that distributions with similar semantics are clustered together. The shapes of images and related descriptions are similar, and the ellipses are closed, showing that the images and text cover similar meanings. For example, since image "4" is a part of image "8", ellipse "8" almost includes all regions of ellipse "4". The intersection of ellipses (images "0", "4", "8" and their corresponding captions) might indicate "a young boy" in images and text. Similar behaviors of our MAP can be found in more visualizations in Appendix B.8. Intuitively, as shown in visualization results, uncertainty modeling facilitates the model to express rich semantic information and complex relationships.

Cases for diverse predictions. Semantic uncertainty is ubiquitous in multimodal tasks. For multimodal understanding tasks such as VQA, an advantage of uncertainty modeling is that multiple predictions can be sampled from distribution representations, which provides diversity. Consider case 3 in Fig. 5, MAP can learn multiple plausible answers (field, park and grass) from the distribution representations, which is close to our real world. In contrast, the point representations from MAP without PDE always generate one answer ignoring other possible expressions. Moreover, distribution representations can also help other multimodal tasks, such as image captioning, to generate several suitable captions, which benefit from diverse correspondences caused by uncertainty modeling.

[1]The images and related captions come from MSCOCO dataset [30].



Case 1
Question: What is this food?
MAP w/o PDE: Cake
MAP:
Dessert
Cake
Muffin

Case 2
Question: What is in the man's ear?
MAP w/o PDE: Headphones
MAP:
Headphones
Headset
Earbuds

Case 3
Question: Where are the people flying kites?
MAP w/o PDE: Field
MAP:
Field
Park
Grass

Figure 5. Predictions sampled from the distribution representations[3].

5. Conclusions

In this work, we focus on the multimodal uncertainty in real-world objects by modeling this onto probability distributions. By considering sequence-level and feature-level interactions, we proposed a Probability Distribution Encoder (PDE) to gain distribution representations for different modalities. Our experiments showed that distribution representations are beneficial for the VL downstream tasks. In addition, uncertainty modeling facilitates diverse predictions. To learn multimodal uncertainty in large-scale data, we designed three new pre-training tasks (D-MLM, D-ITM and D-VLC). Furthermore, we propose an end-to-end **M**ultimodal uncertainty-**A**ware vision-language **P**re-training model (MAP) to obtain generic distribution representations. We demonstrate the effectiveness of the proposed MAP on several VL downstream tasks empirically. In the future, we will explore more distribution subspaces and experiments on larger datasets.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61991450) and the Shenzhen Science and Technology Program (JSGG20220831093004008; ZDSYS20200811142605016).

References

[1] Haifa Alwahaby, Mutlu Cukurova, Zacharoula Papamitsiou, and Michail Giannakos. The evidence of impact and ethical considerations of multimodal learning analytics: A systematic literature review. *The Multimodal Learning Analytics Handbook*, pages 289–325, 2022. 14

[2] Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. In *Proc. of ACL*, 2017. 2

[3] Shruti Bhargava and David A. Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *CoRR*, abs/1912.00578, 2019. 14

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proc. of CVPR*, 2020. 2

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 11

[6] 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 *Proc. of ECCV*, 2020. 2, 4, 6, 12, 14

[7] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proc. of CVPR*, 2021. 1, 2, 7, 13

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019. 2, 11, 12

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Proc. of CVPR*, 2022. 5, 6, 11, 12, 14

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 6, 12

[11] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Proc. of EMNLP Findings*, 2020. 2

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proc. of CVPR*, 2017. 6, 13

[13] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 2019. 1

[14] Eyad Hakami and Davinia Hernández Leo. How are learning analytics considering the societal values of fairness, accountability, transparency and human well-being?: A literature review. *Martínez-Monés A, Álvarez A, Caeiro-Rodríguez M, Dimitriadis Y, editors. LASI-SPAIN 2020: Learning Analytics Summer Institute Spain 2020: Learning Analytics. Time for Adoption?; 2020 Jun 15-16; Valladolid, Spain. Aachen: CEUR; 2020. p. 121-41*, 2020. 14

[15] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez Opazo, and Stephen Gould. VLN BERT: A recurrent vision-and-language BERT for navigation. In *Proc. of CVPR*, 2021. 3

[16] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proc. of CVPR*, 2021. 3

[17] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 11

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. of ICML*, 2021. 2, 3, 11

[19] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proc. of AAAI*, 2020. 2

[20] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. In *Proc. of ICCV*, 2021. 3

[21] Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 1960. 4

[22] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 2006. 4

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proc. of ICML*, 2021. 2, 3, 6, 11, 12

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. of ICLR*, 2014. 5

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 2017. 5, 11

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Proc. of NeurIPS*, 2021. 3, 4, 6, 12

[27] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proc. of ACL*, 2021. 6, 12

[28] Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the geometry of probabilistic box embeddings. In *Proc. of ICLR*, 2019. 2

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. of ECCV*, 2020. 6, 12

[30] 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 *Proc. of ECCV*, 2014. 1, 5, 6, 8, 11, 13, 14

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019. 4

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proc. of NeurIPS*, 2019. 4, 14

[33] Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. *Proc. of NeurIPS*, 2017. 4

[34] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 5, 11

[35] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 6, 11, 13

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, 2021. 2, 3, 4

[37] Tetsuya Sakai. *Laboratory experiments in information retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. Springer, 2018. 6, 13

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL (1)*, 2018. 5, 11

[39] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *Proc. of ICLR*, 2020. 4

[40] Yukun Su, Guosheng Lin, Ruizhou Sun, Yun Hao, and Qingyao Wu. Modeling the uncertainty for self-supervised 3d skeleton action representation learning. In *Proc. of ACM MM*, 2021. 2

[41] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proc. of ACL*, 2019. 6, 13

[42] Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *Proc. of ECCV*, 2020. 2

[43] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *Proc. of EMNLP*, 2019. 4

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of NeurIPS*, 2017. 2, 11

[45] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *Proc. of ICLR*, 2015. 2

[46] Junjie Wang, Yatai Ji, Jiaqi Sun, Yujiu Yang, and Tetsuya Sakai. Mirtt: Learning multimodal interaction representations from trilinear transformers for visual question answering. In *Proc. of EMNLP Findings*, 2021. 2

[47] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *CoRR*, 2021. 6, 12

[48] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, 2021. 6, 12

[49] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *CoRR*, 2018. 6, 13

[50] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning. In *Proc. of ACL*, 2021. 2

[51] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proc. of CVPR*, 2021. 1, 2, 7

[52] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proc. of CVPR*, 2021. 2

[53] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *CVPR*, pages 15650–15659. IEEE, 2022. 6, 12

[54] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *Proc. of ICCV*, 2019. 2, 7

[55] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proc. of CVPR*, 2019. 14

[56] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proc. of CVPR*, 2021. 4, 6, 12

Appendix

A. Architecture Details

A.1. Multi-head Attention

By following the instruction from Transformer [44], the Q, K, V are computed from input hidden states $H \in \mathbb{R}^{T\times D}$ and $H' \in \mathbb{R}^{T'\times D}$. The two input matrices consist of respectively T and T' tokens of d dimensions each. The transformation is as follows:

$$\begin{aligned} Q &= HW_Q \quad W_Q \in \mathbb{R}^{D\times d_k}, \\ K &= H'W_K \quad W_K \in \mathbb{R}^{D\times d_k}, \\ V &= H'W_V \quad W_V \in \mathbb{R}^{D\times d_k}. \end{aligned} \tag{15}$$

An attention map is computed by the pairwise similarity between two tokens from H and H'.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(QK^\top/\sqrt{d_k}\right)V, \tag{16}$$

After splitting k heads from H and H', the Multi-Head Attention (MHA) is concatenated from the outputs by running k attention operations. The same calculations of Q, K, V are conducted in each $i \in [k]$ head to form $Q^{(i)}$, $K^{(i)}$, $V^{(i)}$.

$$\begin{aligned} Head^{(i)} &= \text{Attention}(Q^{(i)}, K^{(i)}, V^{(i)}), \\ \text{MHA}(Q,K,V) &= \text{concat}_{i\in[k]}\left[Head^{(i)}\right] W_O, \end{aligned} \tag{17}$$

where the weight $W_O \in \mathbb{R}^{kd_k\times D}$ projects the concatenation of k head results to the output space D with the same dimension of the inputs. In our models, we set $d_k = D/k$. The other contents in the transformer block, such as MLP Block and residual connection, follow the instructions of Transformer [44]. D is set to 768 and k is set to 12 in our experiments. The detailed experiment settings are presented in the open-sourced codes.

A.2. Multi-head operation in PDE

The operation is similar to the aforementioned multi-head attention in A.1. In this operation, the input hidden states $H \in \mathbb{R}^{T\times D}$ are split into k heads, where T is sequence length and D is hidden size. In each head, we split the features and send them to two paths (μ, σ^2). The operation in the σ^2 path is followed:

$$\begin{aligned} &[Q^{(i)}_{\sigma^2}, K^{(i)}_{\sigma^2}, V^{(i)}_{\sigma^2}] = HW_{qkv}, \\ &Head^{(i)}_{\sigma^2} = \text{Act}\left(Q^{(i)}_{\sigma^2}{K^{(i)}_{\sigma^2}}^\top/\sqrt{d_k}\right)V^{(i)}_{\sigma^2}, \\ &\text{MH}_{\sigma^2}(Q^{(i)}_{\sigma^2}, K^{(i)}_{\sigma^2}, V^{(i)}_{\sigma^2}) = \text{concat}_{i\in[k]}\left[Head^{(i)}_{\sigma^2}\right] W_O, \end{aligned} \tag{18}$$

where d_k is set to $D/(2k)$. The weight $W_{qkv} \in \mathbb{R}^{d_k\times 3d_k}$ projects the inputs to the sub-space in each head. The weight $W_O \in \mathbb{R}^{kd_k\times D}$ projects the concatenation of k head results to the output space. The "Act" is an activation function and normalization function for considering sequence-level interaction. Moreover, the "MH" is the multi-head operation. On the σ^2 path, since the predicted vector has negative values from the activation function, PDE is expected to predict $\log\sigma$. After a simple exp operation, variance vectors are obtained. There are some candidate activation functions are considered: ReLU, ReLU^2, Sigmoid, and Softmax. Unless otherwise specified, the function Softmax is employed in PDE.

Dataset	#Images	#Text
Flickr30K [35]	29K	145K
GQA [17]	79K	1M
MSCOCO [30]	113K	567K
VG [25]	108K	5.4M
SBU [34]	875K	875K
CC-3M [38]	3.1M	3.1M
CC-12M [5]	12M	12M
ALIGN [18]	1.8B	1.8B

Table 7. Details of pre-training datasets in Table 8.

A.3. D-MLM settings

Masked Language Modeling (MLM) is first utilized as a pre-training strategy of BERT [8] to predict masked words, which enhances the ability of contextual modeling. In multimodal pre-training, the missing words are reconstructed with retained text and information from another modality. The model can correctly identify the entity relationships between text and images, learning cross-model semantic alignment. Following the settings from several multimodal models [9, 23], the model randomly covers the text tokens with a probability of 15%, where 80% tokens are replaced with [MASK] token, 10% tokens are replaced with other random words, and 10% tokens remain unchanged.

B. Experiment Details

B.1. Experimental settings

Our experiments are conducted on 8 NVIDIA A100 GPUs. For usual settings in all experiments, we adopt the AdamW optimizer. The learning rate is warmed up first and then decayed linearly. When sampling point vectors from distribution representations, the sample number K is set to 5. In the pre-training phase, the model is trained for 100K steps with a batch size of $4,096$. The learning rate of feature extractors is set to $1e-5$. Cross-modal transformer and PDE's learning rates are both $5e-5$.

For pre-training details, We pre-train our model with D-MLM, D-ITM and D-VLC. In Equation 5 of D-VLC, a

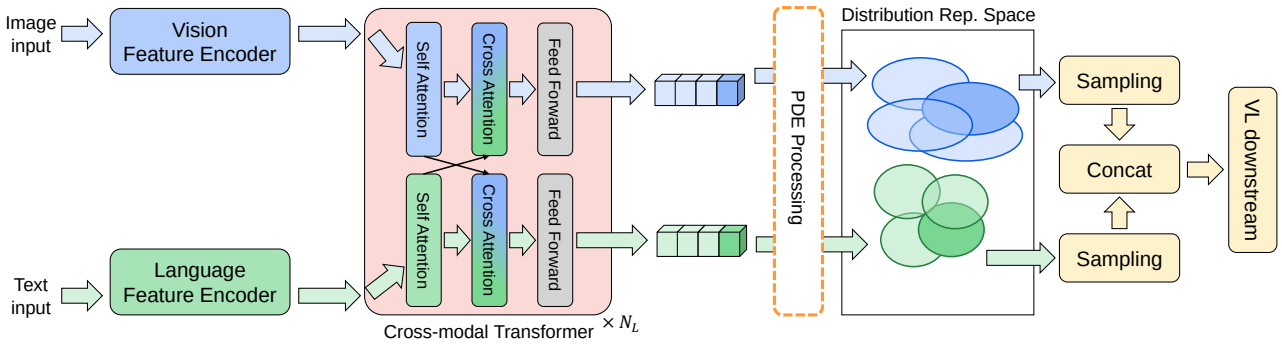


Figure 6. Fine-tuning MAP on different VL downstream tasks.

Model	Paper	Pre-training Datasets	Model size
Pre-training datasets include $>$ 10*M images*			
ALBEF (14M)	[26]	MSCOCO, VG, CC-3M, SBU, CC-12M	Base
SimVLM-base	[48]	ALIGN	Base
Pre-training datasets include $<$ 10*M images*			
UNITER-Large	[6]	MSCOCO, VG, CC-3M, SBU	Large
VILLA-Large	[10]	MSCOCO, VG, CC-3M, SBU	Large
UNIMO-Large	[27]	MSCOCO, VG, CC-3M, SBU	Large
VinVL-large	[56]	MSCOCO, CC-3M, SBU, F30k, GQA	Large
ViLT	[23]	MSCOCO, VG, CC-3M, SBU	Base
UNITER -Base	[6]	MSCOCO, VG, CC-3M, SBU	Base
OSCAR-Base	[29]	MSCOCO, VG, CC-3M, SBU	Base
UNIMO-Base	[27]	MSCOCO, VG, CC-3M, SBU	Base
ALBEF (4M)	[26]	MSCOCO, VG, CC-3M, SBU	Base
VLMo-Base	[47]	MSCOCO, VG, CC-3M, SBU	Base
TCL	[53]	MSCOCO, VG, CC-3M, SBU	Base
METER	[9]	MSCOCO, VG, CC-3M, SBU	Base

Table 8. Details of all models in Table 1 and 2.

is set to -0.005 and b is set to 6. In the full loss formula Equation 14, α is equal to 0.01. For the regularization loss of distributions in Eq. 10, the threshold $\gamma = 300$.

Table 7 reports the statistic of images and text of the pre-training datasets in Table 8, which includes the pre-training datasets of all referenced models. Those datasets are constructed by combining public datasets. However, a substantial portion of the image URLs in datasets might be inaccessible now, which makes the number of images less than the statistic.

B.2. Fine-tuning details

The illustration of fine-tuning MAP on the VL downstream tasks is shown in Figure 6. For different downstream tasks, we just design a simple classifier for understanding tasks. We first sample the point vectors from distribution representations of [CLS]. Then we concatenate point representations from different modalities as global features to conduct classification and apply average pooling operation to all samples' results. The model MAP is trained for 10 epochs. The learning rates of feature extractors, Cross-modal transformer and PDE are $5e-6$, $2.5e-5$, and $2e-4$. In future work, we would like to try applying MAP to do several generation tasks by designing a simple decoder.

B.3. Comparison details

We summarize all referenced modes with model size and pre-training datasets in Table 8. The reported scores in Table 1 and 2 come from their papers. As described in Section 4.2, we introduce the definition of model size [56]. In detail, considering model parameter efficiency, the model size of Vision Language Pre-training (VLP) models can be categorized into at least 3 size: Small, Base, and Large. (1) "Small" indicates the small models prior to the transformer-based VLP models. (2) "Base" indicates the VLP models with similar size to BERT-Base [8]. (3) "Large" is the VLP

Run1	Run2	VQA2.0	SNLI-VE	NLVR2
rand_point	rand_MAP	$p < 0.001$ (-0.193)	$p < 0.001$ (-0.183)	$p < 0.001$ (0.267)
rand_point	pt_point	$p < 0.001$ (-0.211)	$p < 0.001$ (0.028)	$p < 0.001$ (-0.017)
rand_point	pt_MAP	$p < 0.001$ (-0.052)	$p < 0.001$ (0.098)	$p < 0.001$ (0.367)
rand_MAP	pt_point	$p < 0.001$ (-0.018)	$p < 0.001$ (0.211)	$p < 0.001$ (-0.284)
rand_MAP	pt_MAP	$p < 0.001$ (0.141)	$p < 0.001$ (0.280)	$p < 0.001$ (0.100)
pt_point	pt_MAP	$p < 0.001$ (0.159)	$p < 0.001$ (0.070)	$p < 0.001$ (0.384)

Table 9. Statistical significance calculated by Randomized Tukey HSD tests for Table 3 after 1,000 trials. p-value and (effect size) for different tasks.

	MLP	ReLU	ReLU^2	Sigmod
ReLU	$p < 0.001$ (0.385)	-	-	-
ReLU^2	$p < 0.001$ (0.287)	$p < 0.001$ (-0.098)	-	-
Sigmod	$p < 0.001$ (0.162)	$p < 0.001$ (-0.223)	$p < 0.001$ (-0.125)	-
PDE	$p < 0.001$ (-0.151)	$p < 0.001$ (0.234)	$p < 0.001$ (0.136)	$p < 0.001$ (0.011)

Table 10. Statistical significance calculated by Randomized Tukey HSD tests for Table 4 after 1,000 trials. p-value and (effect size).

model with a similar size to BERT-Large. Furthermore, the details of pre-training datasets are presented in Table 7.

B.4. VL downstream tasks

B.4.1 Visual Question Answering

Given an image and a corresponding question, VQA2.0 [12] is the task of providing a correct answer to the question.

B.4.2 NLVR2

The NLVR2 [41] task requires the system to judge whether the corresponding relationship between the description and two images is consistent.

B.4.3 SNLI-VE

SNLI-VE [49] task requires understanding three categories of relationships between images and text, which are entailment, neutral or contradiction.

B.4.4 Image-Text Retrieval

MSCOCO [30] and Filkr30K [35] includes two tasks: Image-to-Text retrieval task and Text-to-Image retrieval task. Both tasks require the model to rank the images or text by computing the image-text similarity scores. In detail, we utilize the Karpathy & Fei-Fei 5K MSCOCO test set and Filkr30K test set and then report the top-K retrieval results.

B.5. Additional results for random initialized MAP

To examine the effectiveness of MAP without extra data, we compare MAP on the popular VL understanding task VQA2.0, with the existing reported methods. As shown in Table 13, MAP achieves a SOTA performance on VQA2.0 among the existing methods without extra data. It shows that PDE can bring multimodal uncertainty knowledge to the models without transferring from large-scale pre-training datasets.

B.6. Comparison between MAP and PCME

PCME [7] is a dual-tower architecture for retrieval, which uses soft contrastive loss with sampled points from distributions. In contrast, Our contrastive loss is based on 2W distance, which directly measures multiple distributions. From a quantization perspective, thanks to pre-training, MAP has a significant boost over PCME. On COCO5k test set, PCME's scores are 44.2/31.9 on I2T/T2I, MAP's scores are 79.3/60.9.

B.7. P-value based on Randomized Tukey HSD tests

In all experiments for our implemented models, p-values were obtained using the randomized Tukey HSD test [37]. The names of runs refer to the related tables. In the experiments, we evaluate the test split in all tasks. Table 10 reports the Randomized Tukey HSD tests for Table 4. In details of Table 9, the name of runs follows the rule: W_M, where W $\in \{\text{rand}, \text{pt}\}$ is random initialization or pre-training and M $\in \{\text{point}, \text{MAP}\}$ is utilizing "MAP w/o PDE" or "MAP". Similarly, Table 11 is the conducted tests for Table 6 with name rules: W_L, where W $\in \{\text{rand}, \text{pt}\}$ and L $\in \{2, 4, 6, 8\}$ is the number of layers of cross-modal transformer in MAP. The tests for Table 5 are shown in Table 12.

B.8. Details and additional examples of visualization

After exacting the distribution representations from PDE, we conduct several 2D toy experiments by using clustering

	rand_2	rand_4	rand_6	rand_8	pt_2	pt_4	pt_6
rand_4	$p < 0.001$ (-0.103)	-	-	-	-	-	-
rand_6	$p < 0.001$ (-0.134)	$p < 0.001$ (-0.031)	-	-	-	-	-
rand_8	$p < 0.001$ (-0.150)	$p < 0.001$ (-0.047)	$p < 0.001$ (-0.016)	-	-	-	-
pt_2	$p < 0.001$ (-0.007)	$p < 0.001$ (0.035)	$p < 0.001$ (0.066)	$p < 0.001$ (0.082)	-	-	-
pt_4	$p = 0.005$ (0.004)	$p < 0.001$ (0.107)	$p < 0.001$ (0.138)	$p < 0.001$ (0.154)	$p < 0.001$ (0.072)	-	-
pt_6	$p < 0.001$ (0.059)	$p < 0.001$ (0.161)	$p < 0.001$ (0.192)	$p < 0.001$ (0.208)	$p < 0.001$ (0.126)	$p < 0.001$ (0.054)	-
pt_8	$p < 0.001$ (0.070)	$p < 0.001$ (0.173)	$p < 0.001$ (0.204)	$p < 0.001$ (0.220)	$p < 0.001$ (0.138)	$p < 0.001$ (0.066)	$p < 0.001$ (0.012)

Table 11. Statistical significance calculated by Randomized Tukey HSD tests for Table 6 after 1,000 trials. p-value and (effect size).

Run1	Run2	VQA2.0	SNLI-VE	NLVR2
rand	MLM_ITM	$p < 0.001$ (0.142)	$p < 0.001$ (0.301)	$p < 0.001$ (0.099)
rand	MLM_VLC	$p < 0.001$ (0.149)	$p < 0.001$ (0.264)	$p < 0.001$ (0.047)
rand	ITM_VLC	$p < 0.001$ (-0.624)	$p < 0.001$ (0.588)	$p < 0.001$ (0.122)
rand	MLM_ITM_VLC	$p < 0.001$ (0.195)	$p < 0.001$ (0.079)	$p < 0.001$ (0.202)
MLM_ITM	MLM_VLC	$p < 0.001$ (0.007)	$p < 0.001$ (-0.038)	$p < 0.001$ (-0.052)
MLM_ITM	ITM_VLC	$p < 0.001$ (-0.765)	$p < 0.001$ (0.286)	$p < 0.001$ (0.023)
MLM_ITM	MLM_ITM_VLC	$p < 0.001$ (0.138)	$p < 0.001$ (-0.222)	$p < 0.001$ (0.103)
MLM_VLC	ITM_VLC	$p < 0.001$ (0.053)	$p < 0.001$ (0.324)	$p < 0.001$ (0.075)
MLM_VLC	MLM_ITM_VLC	$p < 0.001$ (-0.772)	$p < 0.001$ (-0.185)	$p < 0.001$ (0.155)
MLM_VLC	MLM_ITM_VLC	$p < 0.001$ (0.818)	$p < 0.001$ (-0.509)	$p < 0.001$ (0.080)

Table 12. Statistical significance calculated by Randomized Tukey HSD tests for Table 5 after 1,000 trials. p-value and (effect size). MLM, ITM, VLC and rand indicate D-MLM, D-ITM, D-VLC and random initialization respectively.

Model	VQA2.0 (test-dev)
ViLBERT [32]	68.93
MCAN [55]	70.63
UNITER [6]	67.03
METER-swin [9]	72.38
METER-clip-vit [9]	71.75
MAP (ours)	**73.35**

Table 13. Evaluation on VQA2.0 of models with random initialization.

algorithms in machine learning. We utilize the pre-trained MAP with PDE to embed images and text onto distribution representations first. Then, the toy experiments are deployed to find non-linear connections from the input high-dimensional data. In detail, we consider the μ and σ^2 representations in the experiments separately and each experiment calculates more than a thousand image-text pairs. Figures 7, 8, 9 shows several additional visualization examples of the distribution representations in different scenarios[4].

B.9. Visualization between point representations and distribution representations

To explore the differences between representations, we compare our method with ALBEF. For ALBEF (4M), we follow the same method and visualize the features of the same image-sentence pairs (see Fig. 10). Compared to ALBEF, our method takes advantages in capturing rich semantics and concepts in these pairs.

[3] All images and related captions come from MSCOCO dataset [30].

C. Ethical Considerations

Multimodal representation learning is a widely used technique that can have ethical effects. Social bias seems to be rooted in the data due to accumulated biases on the web, such as gender bias in MSCOCO [3]. We believe that our framework could be corrupted, leading to bias concerns, such as having preferences towards certain groups or features. Given that the above problems cover a wide range of issues, such as privacy, fairness, and bias [1, 14], we suggest applying our models to specific contextualization examples. Users should also provide open discussions in their specific research areas and industrial environments.

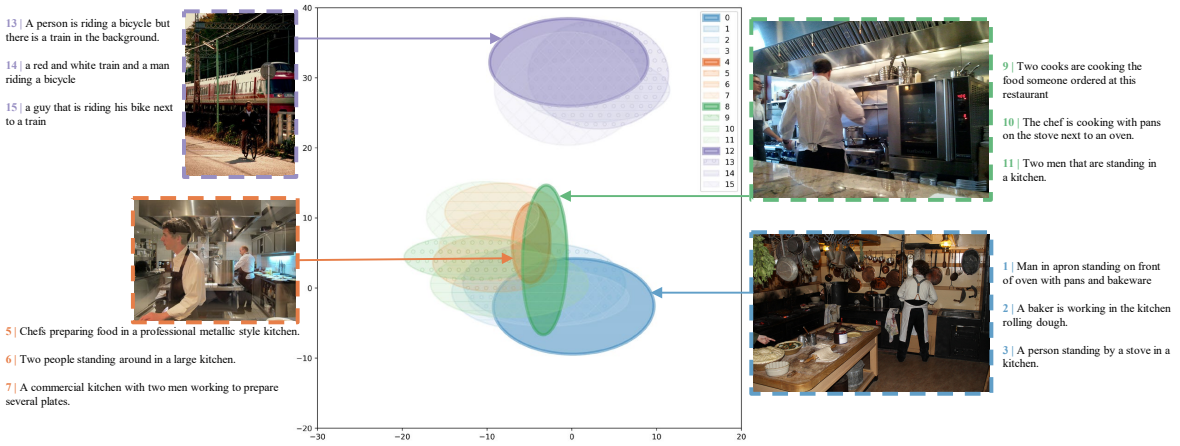


Figure 7. Additional example 1. There are some images and captions of “chef”, “kitchen”, “person”, “bike” and so on.

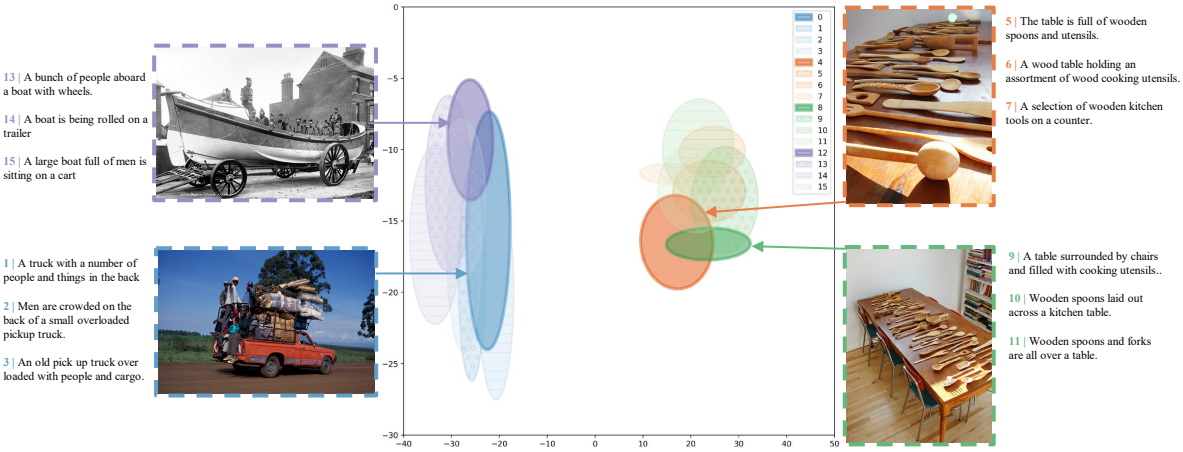


Figure 8. Additional example 2. There are some images and captions of “utensils”, “people”, “truck”, “table” and so on.

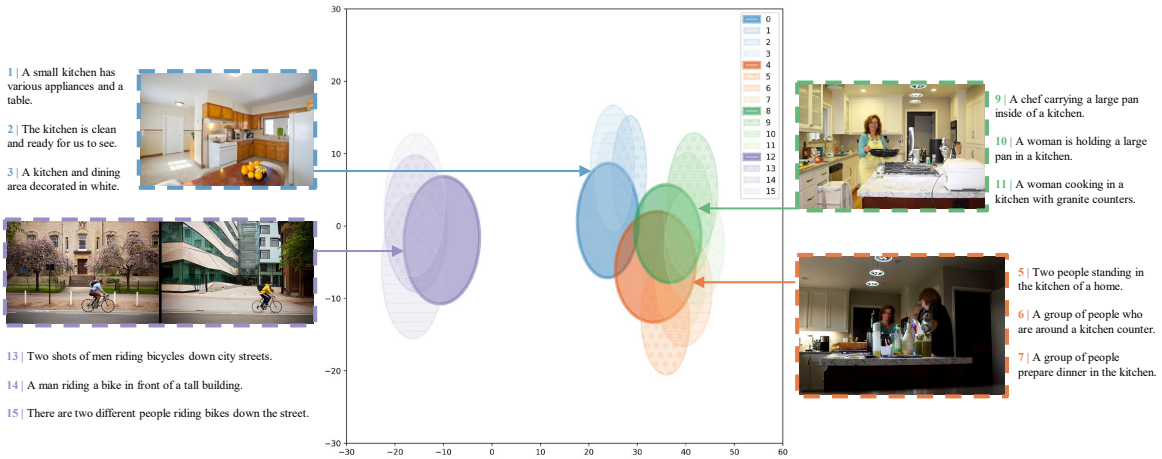


Figure 9. Additional example 3. There are some images and captions of “woman”, “kitchen”, “street”, “bike” and so on.

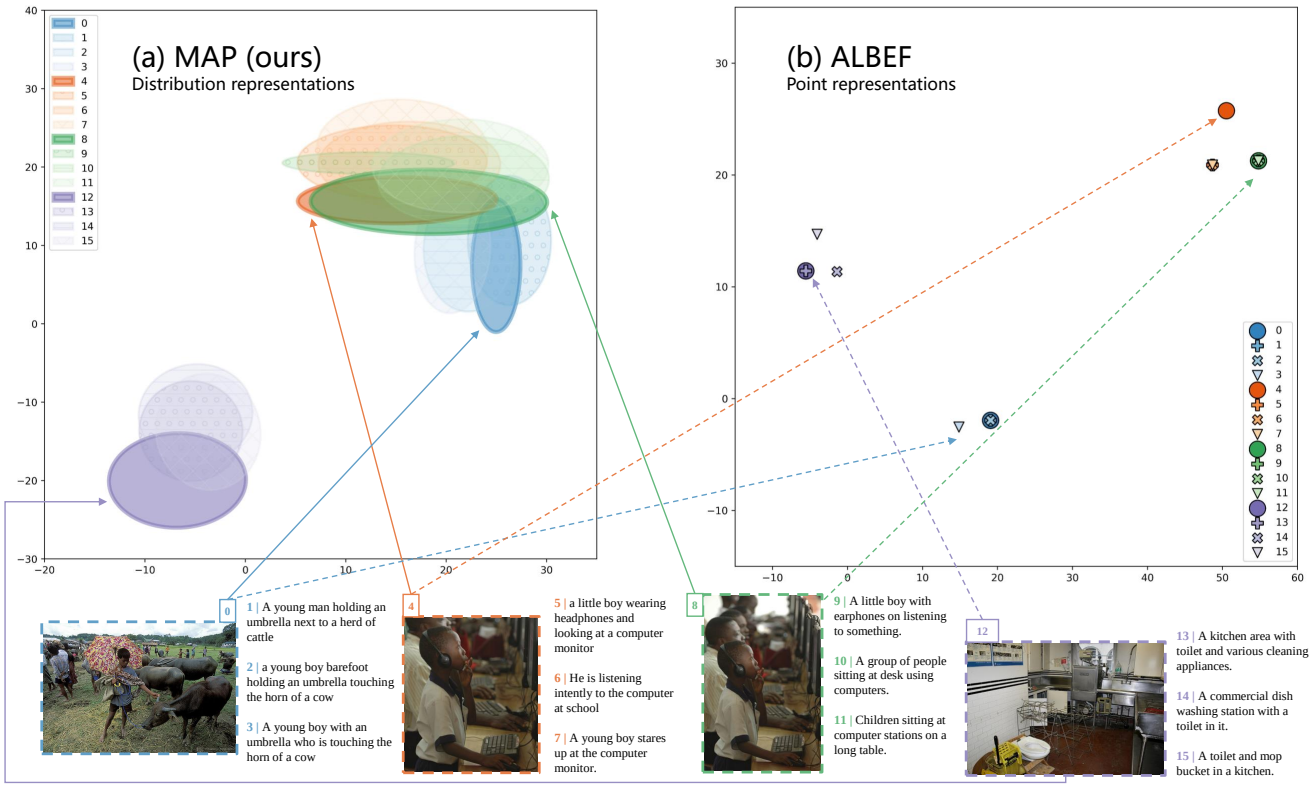


Figure 10. Visualization analysis on distribution representations and point representations.