IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 November 2023, accepted 20 December 2023, date of publication 25 December 2023, date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347192

RESEARCH ARTICLE

Modeling Multimodal Uncertainties via Probability Distribution Encoders Included Vision-Language Models

JUNJIE WANG[1], **YATAI JI**[2], **YUXIANG ZHANG**[1], **YANRU ZHU**[2], **AND TETSUYA SAKAI**[1]
[1]Department of Computer Science and Engineering, Waseda University, Tokyo 169-8050, Japan
[2]Tsinghua University Graduate School, Tsinghua University, Beijing 100190, China

Corresponding author: Junjie Wang (wjj1020181822@toki.waseda.jp)

This work was supported in part by the National Natural Science Foundation of China under Grant 61991450, and in part by the Shenzhen Science and Technology Program under Grant JSGG20220831093004008 and Grant ZDSYS20200811142605016.

ABSTRACT In the field of multimodal understanding and generation, tackling inherent uncertainties is essential for mitigating ambiguous interpretations across multiple targets. We introduce the Probability Distribution Encoder (PDE), a versatile, plug-and-play module that utilizes sequence-level and feature-level interactions to model these uncertainties as probabilistic distributions. Furthermore, we demonstrate its adaptability by seamlessly integrating PDE into established frameworks. Compared to previous methods, our probabilistic approach substantially enriches multimodal semantic understanding. In addition to specific tasks, the unlabeled data contains rich prior knowledge, especially multimodal uncertainties. However, current pre-training methods are designed based on point representations, which hinders the effective functioning of our distribution representations. Therefore, we incorporate this uncertainty modeling into three new pre-training strategies: Distribution-based Vision-Language Contrastive Learning (D-VLC), Distribution-based Masked Language Modeling (D-MLM), and Distribution-based Image-Text Matching (D-ITM). Empirical experiments show that our models achieve State-of-the-Art (SOTA) results in a range of downstream tasks, including image-text retrieval, visual question answering, visual reasoning, visual entailment and video captioning. Furthermore, the qualitative results reveal several superior properties conferred by our methods, such as improved semantic expressiveness over point representations, and the ability to generate diverse yet accurate predictions.

INDEX TERMS Deep learning, modeling uncertainty, multimodal representation learning, pre-training models.

I. INTRODUCTION

Human inherently possesses the ability to comprehend real-word objects with precision, which includes discerning objects with similar semantics and mapping relationships across diverse modalities. Our computational models are engineered to emulate this capability, navigating through complex multimodal semantic landscapes while being acutely aware of inherent data uncertainties. However, the pursuit of such exactitude presents challenges. While multimodal data boasts a rich semantic depth, it also brings about more ambiguity and noise than its single-modality counterparts.

Building upon the pursuit of exactitude in navigating multimodal semantic landscapes, multimodal representation learning techniques serve as a pivotal approach for enhancing sophisticated interpretation across diverse data types [2]. Nevertheless, these methods are not without their own range of challenges. Chief among them is the issue of uncertainty, manifesting both within individual modalities and across different modalities, corroborated by recent works [3], [4]. Consider the image labeled as (a.0) in Fig. 1 as an

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

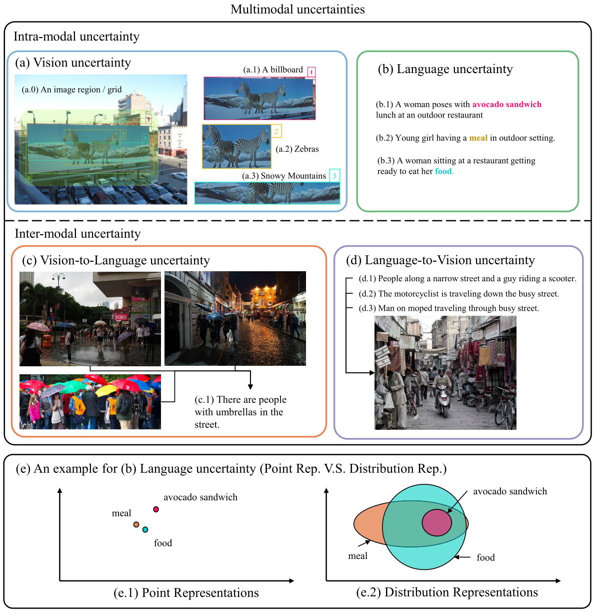
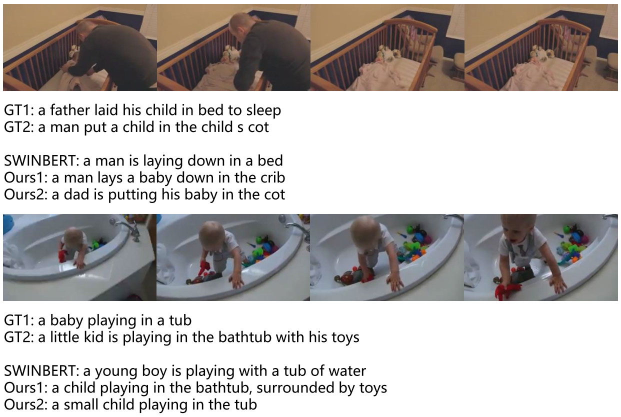

Multimodal uncertainties
Intra-modal uncertainty
(a) Vision uncertainty
(a.0) An image region / grid
(a.1) A billboard
1
(a.2) Zebras
2
(a.3) Snowy Mountains
3
0
(b) Language uncertainty
(b.1) A woman poses with avocado sandwich lunch at an outdoor restaurant
(b.2) Young girl having a meal in outdoor setting.
(b.3) A woman sitting at a restaurant getting ready to eat her food.
Inter-modal uncertainty
(c) Vision-to-Language uncertainty
(c.1) There are people with umbrellas in the street.
(d) Language-to-Vision uncertainty
(d.1) People along a narrow street and a guy riding a scooter.
(d.2) The motorcyclist is traveling down the busy street.
(d.3) Man on moped traveling through busy street.
(e) An example for (b) Language uncertainty (Point Rep. V.S. Distribution Rep.)
avocado sandwich
meal
food
(e.1) Point Representations
avocado sandwich
meal
food
(e.2) Distribution Representations

FIGURE 1. Exemplifying multimodal uncertainties and exploring a case of language uncertainty through point and distribution representations with examples from MSCOCO dataset [1].



GT1: a father laid his child in bed to sleep
GT2: a man put a child in the child s cot

SWINBERT: a man is laying down in a bed
Ours1: a man lays a baby down in the crib
Ours2: a dad is putting his baby in the cot

GT1: a baby playing in a tub
GT2: a little kid is playing in the bathtub with his toys

SWINBERT: a young boy is playing with a tub of water
Ours1: a child playing in the bathtub, surrounded by toys
Ours2: a small child playing in the tub

FIGURE 2. Qualitative examples presenting the effectiveness of our PDE-based framework (SWINPDE) in generating captions. The produced captions maintain semantic coherence and offer a variety of expressions that accurately describe the video content.

example: within a single visual region, one can observe a variety of objects including a billboard, several zebras, and mountains. Consequently, it becomes ambiguous as to which objects are being referred to when discussing this particular region. In the language example labeled as (b) in Fig. 1, complexities arise from intricate relationships among words, contributing to uncertainties like synonymy and hyponymy. In Fig. 1 (c)&(d), the same object is often represented differently across modalities like text and images, exemplifying the challenges of inter-modal uncertainty. The aforementioned multimodal uncertainties pervade a range of tasks in multimodal understanding and generation, such as cross-modal retrieval, visual question-answering, and video captioning. These uncertainties show considerable challenges in the effective training of AI models for these specialized applications. Contrary to addressing these issues, existing methods [5], [6], [7] often overlook these uncertainties, which often results in limited capabilities in comprehending complex concept hierarchies and a lack of prediction diversity. Therefore, it is imperative to *model such multimodal uncertainties.*

In tackling uncertainties inherent to the feature representation space, the utilization of Gaussian distribution stands as a leading approach [3], [8], [9], [10]. In these approaches, the derived uncertainty relies on individual features, neglecting the interplay of all features, which is crucial for understanding inherent relationships. To mitigate this, we employ a specialized component, the Probability Distribution Encoder (PDE), to capture these uncertainty semantics. Beyond the interaction with entire objects, we extend the interactions between word tokens and image patches during the formulation of distribution representations, aiming to learn additional information. In Fig. 1 (e), we showcase two types of representations for language uncertainty, where distribution representations reveal richer semantic relationships than point representations. Moreover, the variance within these distribution representations serves as a metric for text-related uncertainty. Incidentally, distribution representations facilitate diverse generations, yielding several plausible predictions through random sampling. Expanding upon this, in Sec. V-A, we show the effectiveness of the PDE across diverse scenarios of multimodal understanding and generation. Moreover, as illustrated in Fig. 2, we offer a qualitative example in the context of video captioning tasks to provide a deep and intuitive understanding of PDE's functionality. In contrast to the current methods, the PDE module not only contributes to a richer array of possibilities within each video frame but also effectively captures multimodal uncertainties. Consequently, our plug-and-play PDE module serves to enhance the robustness of the models.

In addition to the aforementioned tasks with labeled data, the use of unlabeled multimodal data is flourishing. Concurrent with this trend, various Vision-Language Pre-training (VLP) methods have emerged for self-supervised learning from unlabeled data, offering performance gains in a range of downstream applications [11], [12], [13], [14], [15], [16]. However, existing deterministic representations often lack the ability to grasp uncertainty in pre-training data, as they merely pinpoint positions in semantic space and gauge relationships between targets using certainty metrics like Euclidean distance. *What is the effective approach to modeling multimodal uncertainties within pre-training datasets?*

Therefore, to learn this multimodal uncertainty in a self-supervised manner, we propose a **M**ultimodal uncertainty-**A**ware vision-language **P**re-training (MAP) framework. Specifically, we integrate uncertainty modeling into the multimodal pre-training strategies, yielding the following three new tasks: Distribution-based Masked Language Modeling (D-MLM), Distribution-based Image-Text Matching (D-ITM) and Vision-Language Contrastive learning (D-VLC) pre-training strategies. We construct new objectives and computational processes of these multimodal pre-training tasks, ensuring they effectively adapt to the distribution-based representations. Following fine-grained interactions, D-ITM and D-MLM are deployed for overall-level and token-level alignment of images and text. Moreover, D-VLC addresses coarse-grained multimodal alignment by measuring entire distributions to align representations across modalities.

The contributions of our work are outlined as follows:[1]

- We delve into the multimodal uncertainties. Moreover, we introduce a new plug-and-play module, termed Probability Distribution Encoder (PDE), to model the uncertainty in distribution representations.
- Our proposed PDE methodology enhances the effectiveness of frameworks in various multimodal understanding and generation tasks. Moreover, to the best of our knowledge, we are the first to integrate uncertainty learning into video captioning.
- We formulate three uncertainty-aware multimodal pre-training strategies, namely D-VLC, D-MLM, and D-ITM, to learn multimodal uncertainties in large-scale unlabeled datasets. To our knowledge, this effort represents one of the first attempts to harness the probabilistic nature of distributions in VLP. Our code is available at https://github.com/IIGROUP/MAP.
- We seamlessly integrate our proposed pre-training tasks into a comprehensive end-to-end MAP framework. Moreover, the empirical evaluations demonstrate that the MAP model achieves SOTA performance across multiple downstream tasks. Additionally, our qualitative analysis showcases the effectiveness of our design in capturing multimodal uncertainties within cross-modal tasks, enabling the model to generate diverse and accurate predictions.

II. RELATED WORK

A. PROBABILITY DISTRIBUTION REPRESENTATIONS

Current representation learning approaches predominantly utilize point representations, aiming to closely align these features with the ground truth in high-dimensional space [18], [19]. However, many tasks present inherent uncertainties, suggesting the need for multiple suitable point representations. To tackle this, several works have introduced probability distribution representations, enriching inference and enhancing model robustness to avoid overfitting to a singular solution. Moreover, the recent studies have been conducted to address the uncertainty of input objects, achieving progress in single-modal settings. For instance, W2GM introduces word distributions formed from Gaussian mixtures to cater to multiple word meanings, entailment, and abundant uncertainty information, proposing an energy-based max-margin objective for learning these distributions [20]. Building upon the theme of uncertainty, Smoothed Box employs Gaussian convolutions to craft embeddings under the guidance of uncertain annotations [21]. This approach allows for a nuanced understanding of soft inclusions among various concepts. To tackle the long-tail issue in relation prediction, Gaussian distribution is employed to encapsulate uncertainty in object relationships, aiding scene graph generation [22]. In the multimodal domain, recent efforts in constructing distributions have led to progress in diversifying predictions for cross-modal retrieval tasks [4]. Unlike existing methods that construct distributions at the feature level for an entire image or sentence, our approach models each token within them, like patches in an image and words in a sentence. Consequently, our method is capable of handling interactions at both the sequence-level and feature-level, facilitating the realization of multimodal uncertainty learning.

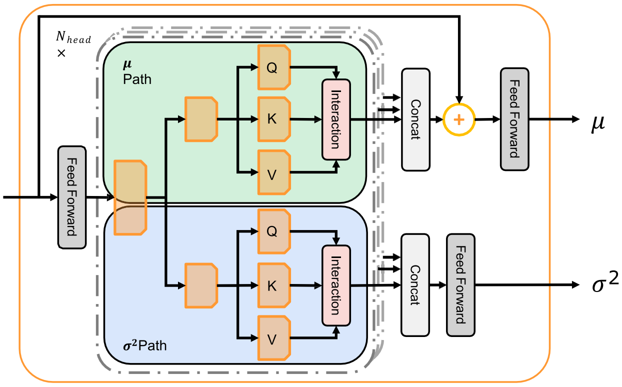


FIGURE 3. Structure of the Probability Distribution Encoder (PDE) module.

B. VISION-LANGUAGE PRE-TRAINING (VLP)

Recently, the Vision-Language (VL) pre-training models have garnered significant attention within the multimodal research community by adeptly addressing real-world challenges through the pre-training and fine-tuning paradigm. In detail, these models initially undergo pre-training tasks on large-scale datasets to acquire common sense knowledge. Following this, they undergo fine-tuning on specific VL downstream tasks, thereby achieving SOTA performance [23], [24]. A core challenge in VLP lies in devising suitable pre-training objectives, with mainstream strategies encompassing Masked Language Modeling (MLM) [15], [25], [26], [27], [28], Image-Text Matching (ITM) [15], [25], [27], [28] and Vision-Language Contrastive (VLC) learning [11], [13], [26], [28]. Specifically, MLM predicts masked tokens using remaining language and vision tokens. ITM assesses the match between different modal inputs, elucidating the alignment between language and vision contexts. VLC discerns inter-modal similarities,

[1] An early version of this paper was presented at CVPR 2023 [17].

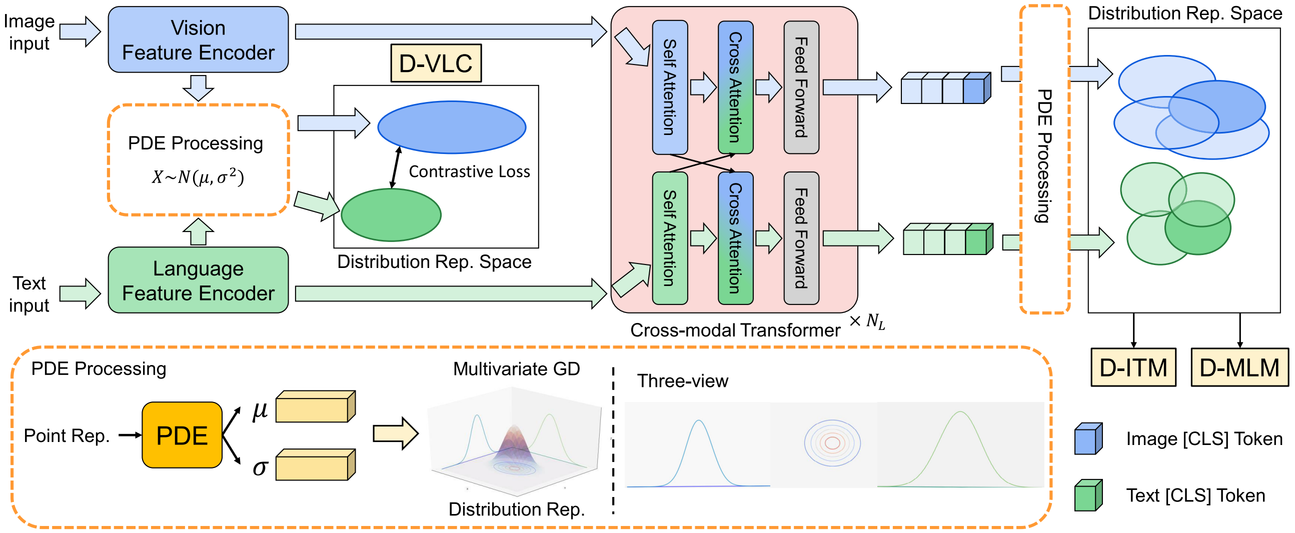


FIGURE 4. MAP's pre-training architecture and objectives: Utilizing PDE to model representations as multivariate Gaussian Distributions (GD). The term "N_L" denotes the cross-modal transformer layer count. And, we provide an illustrated example with a two-dimensional GD. "Rep." indicates representations.

aligning point representations across modalities. Nonetheless, existing methods operate in the point representation space, overlooking the nuances of multimodal uncertainty. To address this, we introduce D-VLC, D-MLM, and D-ITM for pre-training the models within a distribution representation space (Details in Sec. III-C).

III. APPROACHES

Firstly, as shown in Fig. 3, we outline the PDE module in Sec. III-A. Then, we delve into the MAP in Sec. III-B and the overview of it is in Fig. 4. After that, we delineate the distribution-based VLP strategies in Sec. III-C. As detailed in Sec. III-D, following its comprehensive pre-training, the model is fine-tuned on specific VL downstream tasks.

A. PROBABILITY DISTRIBUTION ENCODER (PDE)

The inputs of PDE are derived from the embedding space that encompasses various modalities. To learn the complex multimodal uncertainty, we further model the features using multivariate Gaussian distributions. In detail, for each feature input, PDE calculates a mean vector (μ) and a variance vector (σ^2), where the mean vector represents the central position of distributions in the probabilistic space, and the variance vector depicts the extent of distributions along each dimension. As illustrated in Fig. 3, we present the detailed architecture of PDE, encompassing both sequence-level and feature-level interactions. In particular, the Multi-Head (MH) operation handles sequence-level interactions, whereas the feed-forward layer tackles feature-level interactions. In the MH operation, the input representations $H \in \mathbb{R}^{T \times D}$ are divided into k heads, with T representing the sequence length and D denoting the hidden size. Within each head, the representations are segregated and channeled into two paths (μ, σ^2). In every path, the input representations $H^{(i)} \in \mathbb{R}^{T \times D/2k}$ are projected onto $Q^{(i)}$, $K^{(i)}$, $V^{(i)}$ in the i-th head. For instance, the operation within the μ path is as follows:

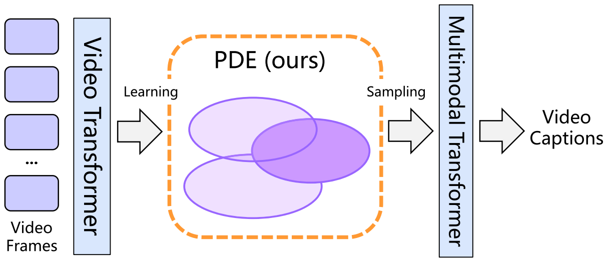


FIGURE 5. The SWINPDE Model integrates PDE module into the SwinBERT [29] architecture for enhanced video captioning.

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

where, d_k is designated the value of $D/(2k)$, and the weight matrix $W_{qkv} \in \mathbb{R}^{d_k \times 3\ d_k}$ aims to project the features into the subspace of each head. Likewise, the weight matrix $W_O \in \mathbb{R}^{kd_k \times D}$ is utilized to project the concatenated results of k heads into the output space. Furthermore, the term "Act" encompasses an activation function and a normalization function, enabling sequence-level interaction. The operations in the σ^2 path are similar to the μ path. Moreover, given the correlation between the input point representation and the mean vector, an "add" operation is utilized to derive the mean vector. The rationale for these design choices is elaborated in Sec. V-C2. Post PDE processing, whether visual or linguistic, each token is depicted as Gaussian distributions within a high-dimensional probabilistic space.

Our design shows that PDE serves as a modular, plug-and-play component, seamlessly integrating with existing point-based frameworks. For instance, to tackle the challenges of video captioning, we incorporate PDE into the widely-used SwinBERT architecture [29], resulting in a modified model known as SWINPDE, as shown in Fig. 5. In this context, we employ a video transformer to extract spatial-temporal video representations from raw video frames. The PDE module then adeptly converts these representations into probabilistic distributions, effectively capturing and learning from the uncertainties. In the final stage, the multimodal transformer processes the sampled representations from the PDE, translating them into coherent natural language sentences for captions via a sequence-to-sequence mechanism. This approach not only enhances the diversity of video captioning but also ensures the adaptability of our module in various scenarios. Furthermore, to extend its applicability to a diverse set of Visual-Language (VL) downstream tasks and pre-training contexts, we embed PDE into our MAP framework, as delineated in Sec. III-B.

B. MODEL OVERVIEW OF MAP

1) FEATURE EXTRACTION

We utilize a vision feature encoder and a language feature encoder. Specifically, the CLIP-ViT [11] serves as the vision encoder, while RoBERTa-Base [30] is employed for language encoding. An image is embedded into a patch feature sequence $\{v_{\texttt{[CLS]}}, v_1, \ldots, v_N\}$, with $v_{\texttt{[CLS]}}$ representing the overall vision feature, and similarly, the input text is transformed into a token sequence $\{w_{\texttt{[CLS]}}, w_1, \ldots, w_M\}$, where $w_{\texttt{[CLS]}}$ denotes the overall language feature.

2) CROSS-MODAL TRANSFORMER

In recent studies, multimodal transformers primarily fall into two categories for fusing diverse modalities: single-stream [12], [31], [32] and dual-stream [28], [33], [34]. In a common setting, the length of image patch sequences significantly surpasses that of text sequences, which poses a challenge for jointly computing attention scores due to the overwhelming weight of vision features [35]. To handle this challenge, we opt for a dual-stream architecture, entailing two separate transformer branches for fusing the input modalities by multiple attention matrices.

As detailed in Fig. 4, we present the overall structure of MAP, including N_L layers of cross-modal encoders. In detail, each encoder layer comprises two Self-Attention (SA) modules and two Cross-Attention (CA) modules. Within the SA block of each modality, the query, key, and value vectors are linearly projected from either vision or language features.

Within the vision-to-language CA module of the i-th layer, the query vectors, embodying language feature T_i' subsequent to the SA module, align with the key or value vectors indicative of vision feature I_i'. The application of the Multi-Head Attention (MHA) operation in the CA block not only enables the integration of visual information across modalities by the language features but also ensures a similar cross-modal interaction in the language-to-vision CA block, mirroring its vision-to-language counterpart. The operations of the i-th layer encoder unfold as follows:

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

C. DISTRIBUTION-BASED PRE-TRAINING TASKS

To capture the uncertainty semantic in common sense, we pre-train the MAP utilizing distribution-based VLP strategies on large-scale unlabeled data. In the pre-training phase, we apply PDEs after feature extractors and cross-modal transformer, respectively. Specifically, the PDE following the feature extractors derives unimodal distribution representations to execute the coarse-grained multimodal alignment. Furthermore, situated at the end of MAP, another PDE is entrusted with fine-grained multimodal alignment.

1) COARSE-GRAINED PRE-TRAINING TASKS

We introduce a method termed **D-VLC** (Distribution-based Vision-Language Contrastive Learning) to achieve coarse-grained multimodal alignment of unimodal representations prior to fusion. Specifically, we employ the 2-Wasserstein distance [36], [37], [38] to measure the distance between multivariate Gaussian distributions. Considering two Gaussian distributions as an example, $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, the 2-Wasserstein distance (D_{2W}) between them is:

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

Furthermore, both Σ_1 and Σ_2 are diagonal matrices, which implies that $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} = \Sigma_1\Sigma_2$. The formula above 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 denotes a standard deviation vector. The distribution representations of the `[CLS]` tokens from the PDEs following feature extractors represent the overall unimodal representations. The similarity between text and an image is calculated as:

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

where a acts as a negative scale factor due to the inverse proportionality of similarity to distance, and b serves as a shift value. Within a batch containing N image-text pairs, we identify N positive matched samples alongside $N(N-1)$

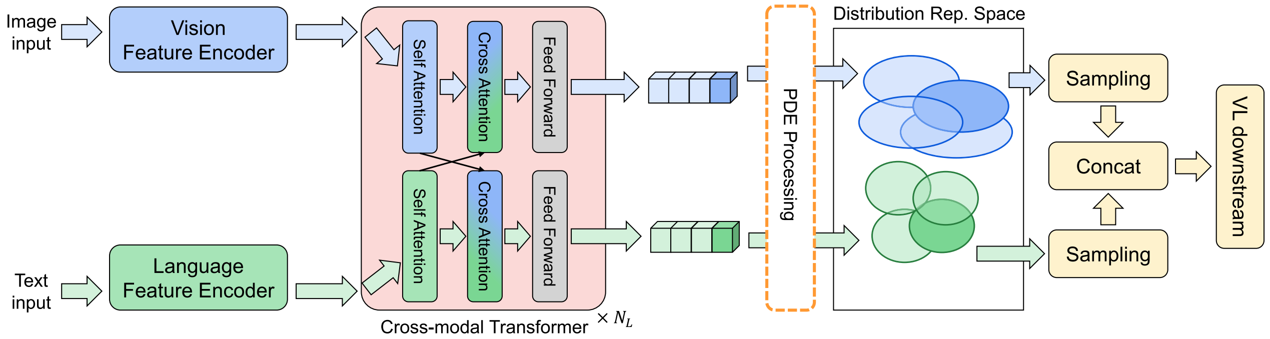


FIGURE 6. Fine-tuning our MAP on the VL downstream tasks. The "N_L" is the layer number of the cross-modal transformer. "Rep." indicates representations.

negative samples, employing the InfoNCE loss as follows:

$$\mathcal{L}^{\text{I2T}}_{\text{InfoNCE}}(i) = -\log \frac{\exp(s(I_i, T_i)/\tau)}{\sum_{n=1}^{N} \exp(s(I_i, T_n)/\tau)},$$
$$\mathcal{L}^{\text{T2I}}_{\text{InfoNCE}}(i) = -\log \frac{\exp(s(T_i, I_i)/\tau)}{\sum_{n=1}^{N} \exp(s(T_i, I_n)/\tau)}, \quad (6)$$

where τ represents a learned temperature parameter. The above expressions are aggregated to form the D-VLC loss $\mathcal{L}_{\text{D-VLC}}$.

2) FINE-GRAINED PRE-TRAINING TASKS

After the cross-modal transformer, fine-grained interaction is enabled on each token across different modalities. We propose two methods to handle the fine-grained multimodal pre-training, which are Distribution-based Masked Language Modeling (D-MLM) and Distribution-based Image Text Matching (D-ITM).

D-MLM necessitates the model to predict masked words by interpreting the text in conjunction with an image. The conventional Masked Language Modeling task, initially employed as a pre-training task for BERT [39], aims at enhancing contextual modeling capabilities. In the VLP scenario, missing words are reconstructed using information from other features and modalities. According to the configurations from several multimodal models [15], [35], the model masks text tokens at a probability of 15%, with 80% of them replaced by the [MASK] token, 10% substituted with random words, and the remaining 10% left unchanged. To predict the masked words, we sample the points from distribution representations, wherein D-MLM minimizes a Cross-Entropy (CE) loss across μ vectors and other sample point vectors:

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

where K denotes the sample number and y represents the label of the masked word. μ is indicative of a mean vector, whereas $z^{(i)}$ stands for stochastic sample point vectors; these vectors are subsequently channeled into the classifier ϕ. In the inference phase, the final output is derived by averaging the prediction results of all samples:

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

D-ITM is a binary classification task that predicts whether a pair of image-text is matched or not. In detail, we extract the point vectors from $w_{\texttt{[CLS]}}$ distributions of multimodal representations, and merge them to generate the results.

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

where v_μ and w_μ represent the mean vectors of vision and language [CLS] distributions, respectively, while $v^{(i)}$ and $w^{(i)}$ denote the sampled points. The D-ITM classifier is denoted by ϕ. The matched image-text pairs serve as positive examples. Negative examples are generated through the random substitution of either images or text descriptions.

3) PRE-TRAINING OBJECTIVES

We observe that random sampling process elevates the training difficulty. Training the model solely with the aforementioned losses induces a variance collapse. As all sampled vectors converge to the optimal position, the distribution eventually degenerates into a point, losing the ability to learn multimodal uncertainty. Hence, we apply a regularization loss to prevent the uncertainty level of distributions from being lower than a specified threshold:

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

where γ serves as a threshold, modulating the uncertainty levels in the learned distributions. Additionally, the function $h(\mathcal{N}(\mu, \sigma^2))$ quantifies the entropy of multivariate Gaussian distributions, as further detailed below:

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

where Σ denotes the covariance matrix, characterized as a diagonal matrix. Furthermore, with the diagonal vector of Σ

TABLE 1. Baseline Model with complexity measured in parameters and data scale in pre-training images.

VLP Model	Paper	Pre-training Datasets	Model Size
Pre-training datasets include $> 10M$ *images*			
ALBEF (14M)	[28]	MSCOCO, VG, CC-3M, SBU, CC-12M	Base
SimVLM-Base	[41]	ALIGN	Base
Pre-training datasets include $< 10M$ *images*			
UNITER-Large	[12]	MSCOCO, VG, CC-3M, SBU	Large
VILLA-Large	[42]	MSCOCO, VG, CC-3M, SBU	Large
UNIMO-Large	[43]	MSCOCO, VG, CC-3M, SBU	Large
VinVL-Large	[32]	MSCOCO, CC-3M, SBU, F30k, GQA	Large
ViLT	[15]	MSCOCO, VG, CC-3M, SBU	Base
UNITER -Base	[12]	MSCOCO, VG, CC-3M, SBU	Base
OSCAR-Base	[44]	MSCOCO, VG, CC-3M, SBU	Base
UNIMO-Base	[43]	MSCOCO, VG, CC-3M, SBU	Base
ALBEF (4M)	[28]	MSCOCO, VG, CC-3M, SBU	Base
VLMo-Base	[45]	MSCOCO, VG, CC-3M, SBU	Base
TCL	[46]	MSCOCO, VG, CC-3M, SBU	Base
METER	[35]	MSCOCO, VG, CC-3M, SBU	Base
MAP	ours	MSCOCO, VG, CC-3M, SBU	Base

being σ^2, (11) can be reformulated as follows:

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

where d indicates the feature dimension.

We observe that the sampling process for $\mathcal{N}(\mu,\sigma^2)$ poses a challenge in inhibiting gradients from propagating back. To address this, by employing the *reparameterization trick* [40], we sample a random variable ϵ from standard normal distributions, rather than directly sampling from $\mathcal{N}(\mu,\sigma^2)$:

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

Following (13), the output z is distributed according to the predicted distributions derived from the PDE. Consequently, we can decouple the calculations of the mean and standard deviation from the sampling operation. This decoupling allows these parameters to be trainable.

In summary, during the pre-training phase, the model executes forward propagation thrice in a single step, conducting the tasks D-MLM, D-ITM, and D-VLC in sequence. Thus, the entire pre-training objective is expressed as follows:

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

where α denotes the weight of $\mathcal{L}_{\text{reg}}$.

D. FINE-TUNING

For applying our MAP model on the VL downstream tasks, we employ the fine-tuning method as illustrated in Fig. 6 after the pre-training stage. To address various downstream tasks, we construct a basic MLP layer for comprehension tasks. Initially, we extract point vectors from the distribution representations of the `[CLS]` tokens. Subsequently, we merge point representations from different modalities as overall features to perform classification, implementing a mean pooling operation on all sampled vectors.

IV. EXPERIMENTAL SETTINGS

A. VL UNDERSTANDING AND GENERATION TASKS

We assess the performance of MAP on the widely recognized Vision-Language (VL) understanding and generation benchmarks, such as video captioning, image-text retrieval, visual question answering, visual reasoning and visual entailment.

1) VIDEO CAPTIONING

In the field of video captioning, MSRVTT [47] stands out with its collection of 10K open-domain video clips, each accompanied by 20 ground-truth captions. Adopting the standard split, our dataset encompasses 6.5K training videos and 2.9K testing videos. We benchmark our method against earlier research using the same test split. In keeping with common evaluation methods [29] in video captioning, we provide detailed comparisons using well-known metrics such as BLEU4, METEOR, ROUGE-L, and CIDEr.

2) IMAGE-TEXT RETRIEVAL

Image-Text retrieval tasks encompass two sub-tasks: Image Retrieval (IR) task and Text Retrieval (TR). Both sub-tasks necessitate the AI system to rank images or text based on the understanding of image-text similarity. We utilize the popular MSCOCO [1] and Flickr30K [48] datasets in the experiments, specifically employing the Karpathy & Fei-Fei

TABLE 2. Details of pre-training datasets in Table 1.

Dataset	# of Images	# of Text
Flickr30K [48]	29K	145K
GQA [52]	79K	1M
CC-12M [53]	12M	12M
ALIGN [13]	1.8B	1.8B
MSCOCO [1]	113K	567K
VG [54]	108K	5.4M
SBU [55]	875K	875K
CC-3M [56]	3.1M	3.1M

5K MSCOCO test set and the Flickr30K test set, and report the top-K retrieval results.

3) VISUAL QUESTION ANSWERING

The objective of the task is to accurately address queries posed in natural language, based on the visual content within provided images. In line with prior work [35], we perform experiments on the VQA2.0 dataset [49], treating the task as a classification problem. For evaluation, we use accuracy as the principal metric.

4) VISUAL REASONING

Within the scope of visual reasoning, the NLVR2 [50] task requires the system to evaluate the consistency between textual descriptions and their corresponding dual-image sets. This dataset encompasses a total of 107, 292 instances, each comprising a human-annotated English sentence paired with two photographs. We utilize accuracy as the metric for evaluation.

5) VISUAL ENTAILMENT

Visual Entailment (VE) is a concept that involves pairs of images and sentences, where the premise is established by an image, diverging from the traditional textual entailment tasks that utilize natural language sentences. The SNLI-VE dataset [51] aims to assess the performance of sophisticated VE models by gauging their ability to accurately infer the semantic congruence between the image and the associated text. Accuracy is employed as the evaluation metric.

B. BASELINES

In our experiments for image-text tasks, our model is compared with an array of SoTA VLP baselines, including but not limited to ALBEF [28] and METER [35]. For a fair comparison environment, we follow the definition of model size [32] for classification. In detail, considering model parameter efficiency, the model size of VLP models can be categorized into at least 2 distinct tiers: Base, and Large. (1) "Base" corresponds to the VLP models with similar size to BERT-Base [39]. (2) "Large" is the VLP model with a similar size to BERT-Large. Furthermore, we summarize all referenced VLP modes with model size and pre-training datasets in Table 1.

TABLE 3. Evaluation on VQA2.0 of models with random initialization. Best scores are in bold.

Model	VQA2.0 (test-dev)
ViLBERT [34]	68.9
MCAN [61]	70.6
UNITER [12]	67.0
METER-swin [35]	72.4
METER-clip-vit [35]	71.8
MAP (ours)	**73.4**

TABLE 4. Evaluation on MSRVTT of models with random initialization. Best results are bolded.

Method	BLEU4	METEOR	ROUGE-L	CIDEr
POS+VCT [57]	42.3	29.7	62.8	49.1
STG-KD [58]	40.5	28.3	60.9	47.1
ORG-TRL [59]	43.6	28.8	62.1	50.9
OpenBook [60]	42.8	29.3	61.7	52.9
SWINBERT [29]	41.9	29.9	62.1	53.8
SWINPDE (ours)	**44.7**	**30.8**	**63.1**	**54.9**

For video-based tasks, we benchmark our approach against existing SOTA methods, such as POS+VCT [57], STG-KD [58], ORG-TRL [59], OpenBook [60], and SWINBERT [29]. In the video-related experiments, we align the same training procedure as the SOTA methods for a fair evaluation environment. In detail, the model parameters of SWINPDE are initialized randomly. Subsequent training occurs on the training set and evaluation is conducted on the test split.

C. PRE-TRAINING DATASETS

Our pre-training datasets comprise MSCOCO [1], Visual Genome (VG) [54], SBU [55] and Conceptual Captions (CC-3M) [56]. Moreover, Table 2 provides statistics on the images and text contained in the pre-training datasets of all referenced models. These datasets are assembled from various public sources. However, some image URLs are unavailable, potentially leading to a lower number of images than initially estimated. During pre-processing phase, we standardize each image into the size of 288 × 288 pixels.

D. IMPLEMENTATION DETAILS

Following a widely-used setting [35], we set the hidden feature sizes to 768, the head number to 12 in the MHA operation, and the layer number (N_L) of the cross-modal transformer to 6. For data processing, each image is resized and cropped to 384 × 384, with the image patch size set to 16. In the term of text, we set the maximum token length to 50. In the PDE module, the default activation function ("Act") in (1) is Softmax and the head number k is set to 6.

TABLE 5. An overall image-text retrieval SoTA comparison with best scores in bold and second best underlined.

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) [28]	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 [12]	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 [42]	-	-	-	-	-	-	76.3	94.2	96.8	87.9	97.5	98.8
UNIMO-Large [43]	-	-	-	-	-	-	78.0	94.2	97.1	89.4	98.9	99.8
VinVL-Large [32]	58.8	83.5	90.3	75.4	92.9	96.2	-	-	-	-	-	-
ViLT [15]	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 [12]	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) [28]	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 [46]	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 [35]	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

In all experiments, the AdamW optimizer is employed, with the learning rate first warmed up and then linearly decayed. In the process of extracting point vectors from distribution representations, we set the sample number K to 5. The experiments are conducted on 8 NVIDIA A100 GPUs.

In the pre-training phase, our MAP is pre-trained with D-MLM, D-ITM, and D-VLC. In detail, a is set to -0.005 and b is set to 6 in (5) of D-VLC task. For the regularization loss of distributions in (10), we set the threshold γ to 300. In (14) of the full loss, α is 0.01. The model undergoes 100K pre-training steps, utilizing a batch size of 4, 096. The learning rate of feature extractors is set to $1e{-}5$, while the cross-modal transformer and the PDE are set to $5e-5$.

In the fine-tuning stage, MAP is trained for 10 epochs, with the learning rates of feature extractors, cross-modal transformer, and PDE set to $5e-6$, $2.5e-5$, and $2e-4$ respectively. In the video-based tasks, SWINPDE is trained on the dataset for 15 epochs, using a learning rate of $3e-4$. Adopting a similar principal architecture to SWINBERT [29], we utilize VidSwin as the encoder for video data and incorporate our PDE to learn the uncertainty. After random sampling from the distribution representations, we engage the same multimodal transformer to decode the captions. For each video in the dataset, random cropping is executed on all frames, consistently targeting the same spatial coordinates to extract a 224 × 224 region.

V. RESULTS AND ANALYSIS

A. RESULTS ON VL UNDERSTANDING AND GENERATION TASKS

1) RESULTS FOR RANDOM INITIALIZED MAP IN VQA2.0

To assess the performance of MAP without the influence of additional data, we compare it with existing methods in the VQA2.0 task for vision-language understanding. Additionally, we aim to understand the effectiveness of the PDE. As revealed in Table 3, MAP performs better than all other methods that do not utilize extra data, achieving SOTA results on the VQA2.0 task. These findings suggest that PDE effectively incorporates multimodal uncertainty into the models, even in the absence of large-scale pre-training datasets. Such outcomes lend further support to the effectiveness and generalizability of our distribution representation modeling approach.

TABLE 6. A comparative analysis with SoTA models on tasks of visual question answering, visual reasoning, and visual entailment is presented. The highest scores are highlighted in bold, while the second highest scores are underlined.

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) [28]	75.84	81.72	81.77	84.20	84.15
SimVLM-Base [41]	77.87	82.55	83.14	80.80	80.91
Group 1: Pre-training datasets include <10M images (Base size)					
ViLT [15]	71.26	75.70	76.13	-	-
UNITER-Base [12]	72.70	77.18	77.85	78.59	78.28
OSCAR-Base [44]	73.16	78.07	78.36	-	-
UNIMO-Base [43]	73.79	-	-	80.00	79.10
ALBEF (4M) [28]	74.54	80.24	80.50	80.14	80.30
VinVL-Base [32]	75.95	82.05	83.08	-	-
VLMo-Base [45]	76.64	82.77	83.34	-	-
METER [35]	77.68	82.33	83.05	80.86	81.19
MAP (ours)	**78.03**	**83.30**	**83.48**	**81.40**	**81.39**

2) RANDOM INITIALIZED PERFORMANCE FOR VIDEO CAPTIONING

Table 4 offers a comprehensive comparison of performance metrics on the MSRVTT datasets, where SWINPDE leads among competitive models. Specifically, SWINPDE marks a great improvement (+2.8) in BLEU4 scores over SWIN-BERT. This outcome shows that our PDE, through the use of distribution representation for visual information, successfully incorporates uncertainty factors, leading to the

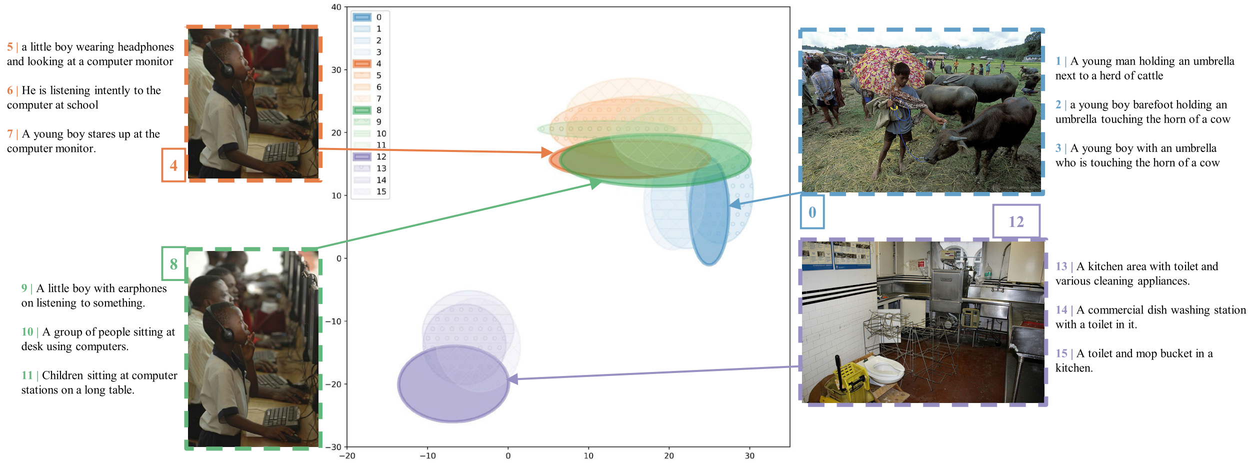


FIGURE 7. Visualization of the distribution representations from pre-trained MAP. The images and related captions come from the MSCOCO dataset. Each 2D Gaussian distribution is represented as an ellipse with 95% confidence. The labels of images and related captions are in the legend.

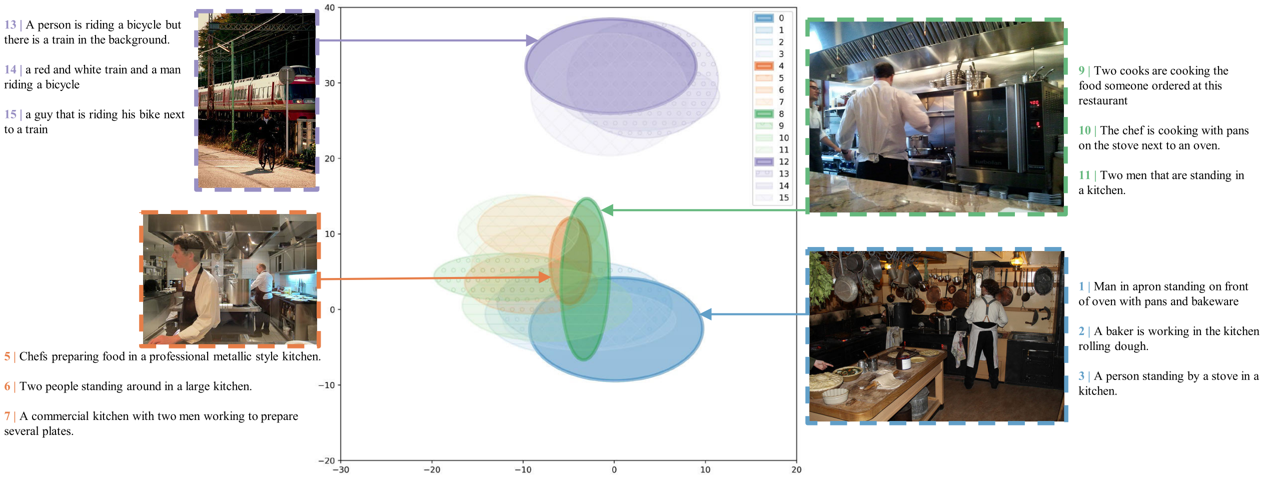


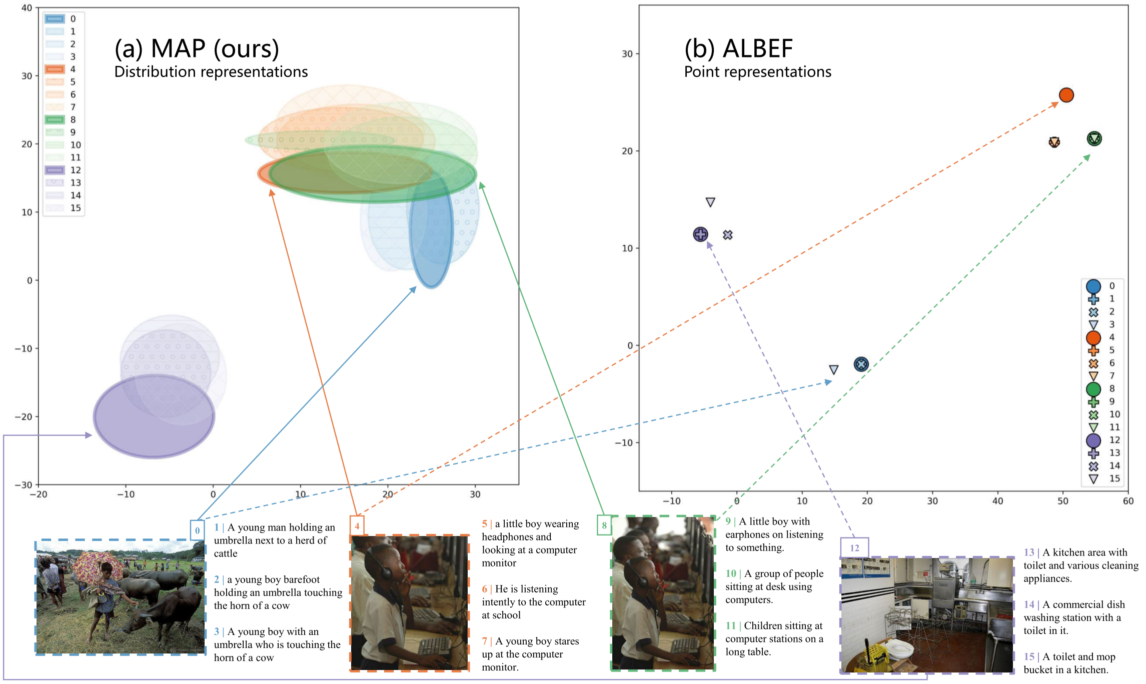
FIGURE 8. An additional example with some images and captions of "chef", "kitchen", "person", "bike" and so on. The samples are from the MSCOCO dataset.

generation of diverse captions. It further validates that our PDE is capable of grasping uncertainty nuances, and distribution representation effectively models this complexity.

3) EVALUATION ON IMAGE-TEXT RETRIEVAL

As shown in Table 5, our MAP model demonstrates superior performance on the MSCOCO dataset, consistently surpassing competing models in all evaluation metrics. On the Flickr30K dataset, the model consistently ranks in either the first or second position across various benchmarks. This is particularly noteworthy, considering that ALBEF, a model specifically designed for retrieval tasks, falls short in comparison to our MAP model. Despite ALBEF's large-scale pre-training with 14M images, our MAP model surpasses it across all metrics on the MSCOCO retrieval task. This performance shows the effectiveness of our approach to uncertainty modeling. On the Flickr30K dataset, our MAP achieves the best performance or only about 0.2 point behind the best score. This slight discrepancy could be attributed to the relatively smaller sample size of Flickr30K compared to other datasets, which might render the model more susceptible to overfitting.

Shifting our focus to a comparison with PCME [4], our MAP model maintains a good performance. Specifically, PCME employs probabilistic distribution representations for retrieval tasks. This superiority is further highlighted on the MSCOCO dataset, where PCME scores 44.2/31.9 on TR@1/IR@1, while our MAP model achieves impressive scores of 79.3/60.9. PCME, which adopts a dual-tower architecture for retrieval, employs a soft contrastive loss with sampled points from distributions. In contrast, our contrastive loss on MAP is based on the 2W distance, a measure that directly handles the multiple distributions.



1 | A young man holding an umbrella next to a herd of cattle

2 | a young boy barefoot holding an umbrella touching the horn of a cow

3 | A young boy with an umbrella who is touching the horn of a cow

5 | a little boy wearing headphones and looking at a computer monitor

6 | He is listening intently to the computer at school

7 | A young boy stares up at the computer monitor.

9 | A little boy with earphones on listening to something.

10 | A group of people sitting at desk using computers.

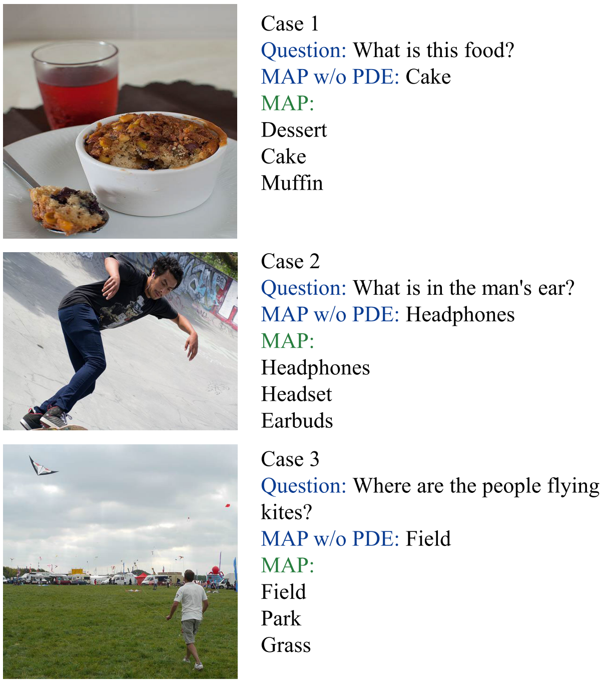
11 | Children sitting at computer stations on a long table.

13 | A kitchen area with toilet and various cleaning appliances.

14 | A commercial dish washing station with a toilet in it.

15 | A toilet and mop bucket in a kitchen.

FIGURE 9. Visualization analysis on distribution representations and point representations. The image-text pairs are from the MSCOCO dataset.



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 10. Predictions sampled from the distribution representations. The samples come from the VQA2.0 dataset.

From a quantization view, pre-training on unlabeled data clearly benefits our MAP model, which outperforms PCME in a great improvement. This comparative analysis further proves the robust performance of our MAP model.

4) EVALUATION ON VQA2.0, NVLR2, AND SNLI-VE

As shown in Table 6, our MAP outperforms the previous SOTA models in Group 1. For example, compared to VLMo-Base, the MAP improves 0.53 points on the NLVR2 dev set. Furthermore, our model achieves a performance boost of +0.35 points on the VQA2.0 test-dev and +0.54 points on the SNLI-VE validation set over METER. It is noteworthy that MAP consistently outperforms SimVLM-Base, which was trained with 1.8 billion pre-training images, across all tasks. This further underscores the effectiveness of our approach to uncertainty modeling.

B. QUALITATIVE RESULTS

Upon deriving the distribution representations from the PDE, we conduct a series of 2D illustrative experiments employing clustering algorithms in Sec. V-B1 and Sec. V-B2. Firstly, we employ the pre-trained MAP to encode images and text into distribution representations. After that, these illustrative experiments are performed to uncover non-linear relationships within the high-dimensional embeddings. Specifically, we separately examine the μ and σ^2 representations in the experiments, with each experiment evaluating over a thousand image-text pairs.

In Sec. V-B3, we examine the impact of PDE on the results generated by MAP in the VQA2.0 task. Specifically, we show the results outputted by the models. For models incorporating PDE, we perform three samplings from the predicted distributions and subsequently convert these embedings back into natural language.

1) VISUALIZATION ON THE DISTRIBUTION REPRESENTATIONS

We focus on the visualization of distribution representations generated by the pre-trained MAP model. Fig. 7 presents the characteristics of distribution representations, showing that distributions with similar semantic meanings tend to cluster. Both the geometry of the images and the textual descriptions manifest congruent characteristics. The completeness of the enclosing ellipses signifies a robust semantic coverage across both visual and textual elements. For example, owing to the part-whole relationship between images "4" and "8", the area under ellipse "8" nearly subsumes that of ellipse "4". The intersection of ellipses for images "0", "4", and "8" and their captions indicate several shared themes (such as "a young boy") in both visual and textual data. Additional similar behavioral patterns of our MAP model are presented in Fig. 8. The qualitative findings affirm that the model's uncertainty modeling serves to encapsulate complex semantic relationships and rich contextual nuances effectively.

2) COMPARATIVE VISUALIZATION OF POINT REPRESENTATIONS AND DISTRIBUTION REPRESENTATIONS

In our study, we aim to explore the differences between the representations generated by our proposed method and those produced by ALBEF [28], a well-known method in the field of multimodal representation learning. As shown in Fig. 9, we follow the same method and visualize the features of the same image-sentence pairs for ALBEF (4M). Compared to ALBEF, our method takes advantage of capturing rich semantics and diverse concepts embedded in the image-sentence pairs. Our method effectively captures intra-modal and inter-modal uncertainty through distribution, reflected in the distribution's characters within the representation space, such as shapes and positions. This capability is crucial for many downstream tasks like visual question answering, requiring a nuanced understanding of both visual and textual content.

3) CASES FOR DIVERSE PREDICTIONS

As illustrated in Fig. 10, we explore the advantages of distribution representation, which allows for a diverse range of prediction results. In the field of multimodal tasks, semantic uncertainty is a prevalent challenge. In multimodal understanding tasks such as VQA, a notable benefit of uncertainty modeling is the ability to sample multiple predictions from distribution representations, thereby yielding diversity. Take Case 3 in Fig. 10 for instance, where MAP furnishes multiple plausible answers (field, park, and grass), closely mirroring real-world scenarios. On the other hand, the point representations from MAP without PDE invariably generate a singular answer, overlooking other possible descriptions. Moreover, the distribution representations extend their utility to other multimodal tasks like video captioning, enabling the generation of diverse and fitting descriptions. This benefit arises from the diversity created by uncertainty modeling.

TABLE 7. 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.

Method	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 8. Effect of different structures of PDE. We explore the different designs of "Act" in Equation 1. Normal denotes the normalization operation.

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**

C. ABLATION STUDIES

1) IMPACT OF PROBABILITY DISTRIBUTION REPRESENTATIONS ON VL TASKS

As illustrated in Table 7, applying PDE improves performance in various VL downstream tasks. Regardless of whether the model is initialized with random or pre-trained weights, distribution representations consistently outperform point representations in terms of VL understanding. The superior performance of distribution representations can be attributed to their ability to capture multimodal uncertainties, thereby conveying a more nuanced semantic understanding.

2) DESIGN CONSIDERATIONS FOR PDE

As illustrated in Table 8, we investigate the influence of various designs on the performance of the PDF. Upon eliminating the sequence-level interaction in PDE, we refer to it as "MLP only" (MultiLayer Perceptron), a prevalent approach utilized in previous studies [3], [4], [9]. Our PDE (Softmax) exceeds the performance of the "MLP only" approach on VQA2.0, thereby gaining an advantage from the sequence-level information. To explore the impact of the structures on the outcomes, we propose several potential activation functions: ReLU, ReLU^2, Sigmoid, and Softmax. The function ReLU indicates the activation status of the relationship between tokens, while ReLU^2 enhances ReLU by being differentiable. We note that "MLP only" surpasses ReLU and ReLU^2, illustrating the importance of sequence-level interaction design. The function Sigmoid maps input data to a range between 0 and 1, smoothly assigning weights among different tokens. Lastly, the function

TABLE 9. The effect of distribution-based pre-training tasks. We pre-train the model on the MSCOCO dataset.

Training strategy	VQA2.0	SNLI-VE	NLVR2
	test-dev	test-p	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 10. The effect of different layer numbers in the cross-modal transformer on VQA2.0.

# of Layers	Random Initializing	Pre-training
2	72.71	73.78
4	73.32	74.73
6	**73.35**	75.16
8	73.31	**75.26**

Softmax surpasses the others in VQA2.0, indicating that Softmax is apt for expressing the correlation between tokens. As a result, we select Softmax as the primary activation function.

3) EVALUATING THE ROLE OF DIFFERENT PRE-TRAINING STRATEGIES

Table 9 shows the influence of different pre-training tasks on the performance of VL downstream tasks. The lack of D-MLM pre-training results in the lowest performance among all pre-training strategies, underscoring the crucial role of D-MLM in pre-training. Additionally, both D-VLC and D-ITM aid the model in understanding the semantic similarity between different modalities. Concerning specific tasks, D-VLC yields more substantial improvements in VQA2.0, while D-ITM proves to be more efficacious in enhancing performance on SNLI-VE and NLVR2.

4) ANALYZING THE EFFICACY ACROSS DIFFERENT LAYER ARCHITECTURES OF THE CROSS-MODAL TRANSFORMER

As shown in Table 10, we explore the influence of layer count in the VQA2.0 task, considering both random initialization and pre-training strategies such as D-MLM, D-ITM, and D-VLC on the MSCOCO dataset. With random initialization, a model with six layers demonstrates optimal performance, albeit hitting a performance plateau. Upon employing pre-training strategies, the eight-layer model surpasses its six-layer counterpart, indicating that pre-training aids in overcoming the bottleneck posed by parameter limitations. This improvement is likely attributed to large-scale data pre-training mitigating the issue of over-fitting when more parameters are involved. Additionally, the benefits of pre-training diminish as the number of layers decreases, owing to the model's constrained learning capacity.

VI. CONCLUSION

In this study, we focus on quantifying the multimodal uncertainties associated with real-world objects via probabilistic modeling. Leveraging both sequence-level and feature-level interactions, we introduce a Probability Distribution Encoder (PDE) designed to obtain distributional representations across various modalities. To facilitate its application, PDE seamlessly integrates into well-established vision-language models, such as SWINPDE. Our qualitative results highlight the advantages of employing distribution representations over point representations, particularly in enhancing semantic expressiveness and diverse predictions upon learning uncertainties. To exploit large-scale unlabeled data for multimodal uncertainty learning, we formulate three new pre-training tasks: D-MLM, D-ITM, and D-VLC. Moreover, we present an end-to-end Multimodal uncertainty-Aware vision-language Pre-training model (MAP) aimed at acquiring generic distributional representations. Empirical evidence suggests that these distribution representations significantly contribute to the performance in vision-language understanding and generation tasks. Our models and methods set new benchmarks, achieving SOTA results on multiple datasets and tasks.

ACKNOWLEDGMENT

(Junjie Wang and Yatai Ji are co-first authors.) The authors would like to thank Yuan Gong, Lin Zhang, Hongfa Wang, Jiaxing Zhang, and Yujiu Yang, for their support and suggestions.

REFERENCES

[1] 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 *Proc. ECCV*, 2014, pp. 740–755.

[2] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[3] G. Yang, J. Zhang, Y. Zhang, B. Wu, and Y. Yang, "Probabilistic modeling of semantic ambiguity for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12522–12531.

[4] S. Chun, S. J. Oh, R. Sampaio de Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8411–8420.

[5] C. Jing, Y. Wu, X. Zhang, Y. Jia, and Q. Wu, "Overcoming language priors in VQA via decomposed linguistic representations," in *Proc. AAAI*, 2020, pp. 11181–11188.

[6] F. Gardères, M. Ziaeefard, B. Abeloos, and F. Lecue, "ConceptBert: Concept-aware representation for visual question answering," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 489–498.

[7] J. Wang, Y. Ji, J. Sun, Y. Yang, and T. Sakai, "MIRTT: Learning multimodal interaction representations from trilinear transformers for visual question answering," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2021, pp. 2280–2292.

[8] L. Vilnis and A. McCallum, "Word representations via Gaussian embedding," in *Proc. ICLR*, 2015.

[9] T. Yu, D. Li, Y. Yang, T. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 552–561.

[10] Y. Su, G. Lin, R. Sun, Y. Hao, and Q. Wu, "Modeling the uncertainty for self-supervised 3D skeleton action representation learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 769–778.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[12] 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 *Proc. ECCV*, 2020, pp. 104–120.

[13] C. Jia, Y. Yang, Y. Xia, Y. 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," in *Proc. ICML*, 2021, pp. 4904–4916.

[14] H. Xu, M. Yan, C. Li, B. Bi, S. Huang, W. Xiao, and F. Huang, "E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning," in *Proc. ACL*, 2021, pp. 503–513.

[15] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. ICML*, 2021, pp. 5583–5594.

[16] Y. Liu, K. Wang, L. Liu, H. Lan, and L. Lin, "TCGL: Temporal contrastive graph for self-supervised video representation learning," *IEEE Trans. Image Process.*, vol. 31, pp. 1978–1993, 2022.

[17] Y. Ji, J. Wang, Y. Gong, L. Zhang, Y. Zhu, H. Wang, J. Zhang, T. Sakai, and Y. Yang, "MAP: Multimodal uncertainty-aware vision-language pre-training model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23262–23271.

[18] Y. Wei, Y. Liu, H. Yan, G. Li, and L. Lin, "Visual causal scene refinement for video question answering," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 377–386.

[19] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11624–11641, Jun. 2023.

[20] B. Athiwaratkun and A. G. Wilson, "Multimodal word distributions," in *Proc. ACL*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 1645–1656.

[21] X. Li, L. Vilnis, D. Zhang, M. Boratko, and A. McCallum, "Smoothing the geometry of probabilistic box embeddings," in *Proc. ICLR*, 2019.

[22] G. Yang, J. Zhang, Y. Zhang, B. Wu, and Y. Yang, "Probabilistic modeling of semantic ambiguity for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12522–12531.

[23] X. Yang, F. Lv, F. Liu, and G. Lin, "Self-training vision language BERTs with a unified conditional model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3560–3569, Jan. 2023.

[24] Y. Ji, R. Tu, J. Jiang, W. Kong, C. Cai, W. Zhao, H. Wang, Y. Yang, and W. Liu, "Seeing what you miss: Vision-language pre-training with semantic completion learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6789–6798.

[25] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "VLN-BERT: A recurrent vision-and-language BERT for navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1643–1653.

[26] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "MDETR-modulated detection for end-to-end multi-modal understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1760–1770.

[27] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12971–12980.

[28] 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," in *Proc. NIPS*, 2021, pp. 9694–9705.

[29] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "SwinBERT: End-to-end transformers with sparse attention for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17928–17937.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[31] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," in *Proc. ICLR*, 2020.

[32] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5575–5584.

[33] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. EMNLP*, 2019, pp. 5099–5110.

[34] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. NIPS*, 2019, pp. 13–23.

[35] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, Z. Liu, and M. Zeng, "An empirical study of training end-to-end vision-and-language transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18145–18155.

[36] A. Mallasto and A. Feragen, "Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes," in *Proc. NIPS*, 2017, pp. 5660–5670.

[37] L. V. Kantorovich, "Mathematical methods of organizing and planning production," *Manage. Sci.*, vol. 6, no. 4, pp. 366–422, Jul. 1960.

[38] L. Kantorovitch, "On the translocation of masses," *Manage. Sci.*, vol. 5, no. 1, pp. 1–4, Oct. 1958.

[39] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014.

[41] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," 2021, *arXiv:2108.10904*.

[42] Z. Gan, Y. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," in *Proc. NIPS*, 2020.

[43] 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," in *Proc. ACL*, 2021, pp. 2592–2607.

[44] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. ECCV*, 2020, pp. 121–137.

[45] W. Wang, H. Bao, L. Dong, and F. Wei, "VLMo: Unified vision-language pre-training with mixture-of-modality-experts," 2021, *arXiv:2111.02358*.

[46] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15650–15659.

[47] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 5288–5296, doi: 10.1109/CVPR.2016.571.

[48] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.

[49] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6325–6334.

[50] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6418–6428.

[51] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment task for visually-grounded language learning," 2018, *arXiv:1811.10582*.

[52] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6693–6702.

[53] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3557–3567.

[54] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[55] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. NIPS*, 2011, pp. 1143–1151.

[56] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

[57] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8917–8926.

[58] B. Pan, H. Cai, D.-A. Huang, K.-H. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, "Spatio-temporal graph for video captioning with knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10867–10876.

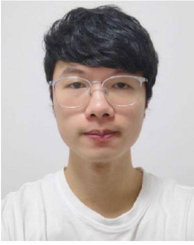
[59] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, and Z.-J. Zha, "Object relational graph with teacher-recommended learning for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13275–13285.

[60] Z. Zhang, Z. Qi, C. Yuan, Y. Shan, B. Li, Y. Deng, and W. Hu, "Open-book video captioning with retrieve-copy-generate network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9832–9841.

[61] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6274–6283.

JUNJIE WANG is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Waseda University. His long-term research goal is to utilize information from different perspectives for building AI/human-AI-interaction systems.

YATAI JI received the bachelor's degree from the Department of Automation, Tsinghua University, China, in 2021. He is currently pursuing the master's degree in electronic and information engineering with the Tsinghua Shenzhen International Graduate School, Tsinghua University. His research interests include deep learning, multi-modal learning, and vision-language pre-training.



YUXIANG ZHANG is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Waseda University.

YANRU ZHU received the master's degree from the Department of Computer and Science. His research interests include multi-modal learning and computer vision.

TETSUYA SAKAI is currently a Professor with the Department of Computer Science and Engineering, Waseda University, Japan. He is also a General Research Advisor with NAVER Corporation, South Korea, and a Visiting Professor with the National Institute of Informatics, Japan. In 2023, he was inducted into the SIGIR Academy. His research interests include information retrieval, natural language processing, and human–computer interaction (HCI).

He is an ACM Distinguished Member and an IPSJ Fellow. He is a Senior Associate Editor of *ACM TOIS*.