Heterogeneous Prompt-Guided Entity Inferring and Distilling for Scene-Text Aware Cross-Modal Retrieval

Zhiqian Zhao[1], **Liang Li**[2*], **Jiehua Zhang**[3], **Yaoqi Sun**[1,5], **Xichun Sheng**[4,5], **Haibing Yin**[1,5], **Shaowei Jiang**[1,5]

[1]Hangzhou Dianzi University, Hangzhou, China,
[2]Institute of Computing Technology, Chinese Academy of Sciences,
[3]School of Software Engineering, Xi'an Jiaotong University,
[4]Macao Polytechnic University, Macao, China,
[5]Lishui Institute of Hangzhou Dianzi University
{zhiqian.zhao, syq, yhb, jiangsw}@hdu.edu.cn, liang.li@ict.ac.cn, jiehua.zhang@stu.xjtu.edu.cn, p2314922@mpu.edu.mo

Abstract

In cross-modal retrieval, comprehensive image understanding is vital while the scene text in images can provide fine-grained information to understand visual semantics. Current methods fail to make full use of scene text. They suffer from the semantic ambiguity of independent scene text and overlook the heterogeneous concepts in image-caption pairs. In this paper, we propose a heterogeneous prompt-guided entity inferring and distilling (HOPID) network to explore the nature connection of scene text in images and captions and learn a property-centric scene text representation. Specifically, we propose to align scene text in images and captions via heterogeneous prompt, which consists of visual and text prompt. For text prompt, we introduce the discriminative entity inferring module to reason key scene text words from captions, while visual prompt highlights the corresponding scene text in images. Furthermore, to secure a robust scene text representation, we design a perceptive entity distilling module that distills the beneficial information of scene text at a fine-grained level. Extensive experiments show that the proposed method significantly outperforms existing approaches on two public cross-modal retrieval benchmarks.

Demo — https://my-hopid.github.io

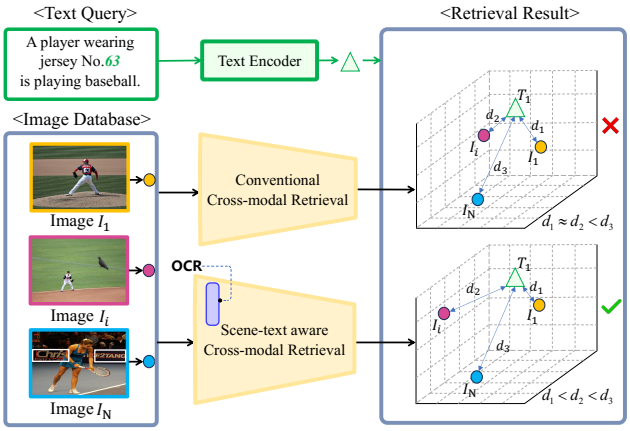

Figure 1: Given a text query, conventional cross-model retrieval cannot distinguish similar images in the common embedding space ($d_1 \approx d_2 < d_3$). Scene-text aware cross-modal retrieval can reduce the distance between matching image and text pair in the embedding space ($d_1 < d_2 < d_3$).

Introduction

Cross-modal retrieval, which involves searching for relevant images in a database by giving a text query or vice-versa, has gained attention due to its wide applications. Significant progress (Lee et al. 2018; Li et al. 2019; Wang et al. 2020a; Zhang et al. 2020b; Koley et al. 2024; Xie et al. 2024) has been made through global alignment of image and text, or local alignment of salient regions and corresponding words. Despite this advance, these object detector-based methods cannot capture the text appearing in images resulting ineffective in text-related scenarios. To address this problem, researchers (Mafla et al. 2021; Cheng et al. 2022; Zhou et al. 2022; Miyawaki et al. 2022; Zhou et al. 2023, 2024) try to explore the scene-text aware cross-modal retrieval for comprehensively understanding visual semantics.

To better leverage scene text, current works can be divided into two categories: local-align methods (Zhou et al. 2022; Miyawaki et al. 2022; Zhou et al. 2023) and global-align methods (Mafla et al. 2021; Cheng et al. 2022). **1) For local-align methods**, Zhou et al. (2022, 2023) propose to align textual OCR in the caption with the visual OCR in the image word by word. Further, recent work (Miyawaki et al. 2022) introduces a co-mask strategy to locally align the scene text in images and captions. Although these approaches are intuitive and interpretable, excessive emphasis on local regions often leads to the sensitivity of local noise. **2) For global-align methods**, they propose to align the caption with the fusion of scene text and other visual clues in the image. In detail, STARNet (Mafla et al. 2021) builds the relationship among scene texts to obtain high-level semantic information of scene text by a graph convolutional network (GCN). Although the relationship between scene texts is important, paying too much attention on the high-level semantic language information without other multi-modalities

*Corresponding Author

information of scene texts may lead to the semantic ambiguity issue. More recently, ViSTA (Cheng et al. 2022) regards scene text as another modality excepting image and caption in the cross-modal retrieval task. ViSTA fuses scene texts with other visual components by a fusion token to aggregate critical visual entities and scene text semantics, then globally aligns it with the caption. This method provides a more efficient global alignment, but this simplistic global matching overlooks the natural heterogeneous concepts of scene text in image-caption pairs, hindering the fine-grained alignment of scene text, and thus limiting the accuracy.

In fact, humans are good at scene-text aware cross-modal retrieval by carefully comparing discriminative entities in images and captions, especially the scene texts, *i.e.*, finding whether there is a shared scene text in both image and caption. Besides, humans can comprehensively understand scene texts in images based on their contextual information and extract valuable entity information from them, which is beneficial for alleviating semantic ambiguity. Inspired by this, we argue that, (1) aligning discriminative scene text in both images and captions is beneficial for mining the inherent connection of heterogeneous scene text. (2) Understanding scene text comprehensively with multi-modal context is conducive to cross-modal retrieval.

In light of this, we propose a novel **H**eter**O**geneous **P**rompt-guided entity **I**nferring and **D**istilling (HOPID) network to reason discriminative multi-modal entities for aligning cross-modal heterogeneous scene text and learn property-centric scene text representation to ease semantic ambiguity. First, we devise a Discriminative Entity Inferring (DEI) module to predict the discriminative scene text words from caption and then integrate them with the pre-defined prompt template to build the text prompt. Meanwhile, for the visual prompt, we explicitly locate scene text in image with a visual clue marker to hint at their position and provide the surroundings as the visual context for alleviating the local noise. The heterogeneous prompt, consisting of text and visual prompt, utilizes the OCR-aware contrastive learning to align the different manifestations of the same entity (scene text in images and captions) to narrow the cross-modal gap. Second, to alleviate the semantic ambiguity of semantic-only scene text, we design a Perceptive Entity Distilling (PED) module that extracts salutary multi-model scene text information in an iterative manner to learn a more robust scene text representation. Specifically, PED utilizes the slot attention mechanism to explore OCR features from various modalities ($i.e.$, visual, semantics, and position) and distill valuable scene text entities with property-centric representation. Therefore, PED interprets scene text from a property perspective and mitigates the semantic ambiguity of isolated scene text. We conduct extensive experiments on two public benchmarks and achieve a new state-of-the-art performance.

Our contributions are summarized as follows: *1)* We propose a new HOPID network to explore the heterogeneous scene text in image and caption and acquire property-centric scene text representation, thus achieving a more comprehensive utilization of scene text in cross-modal retrieval. *2)* We introduce heterogeneous prompt reasoning to mine the inherent connection of scene text from image and caption for alignment. *3)* We introduce a perceptive entity distilling module that extracts salutary multi-modalities information among scene texts to obtain a robust representation for retrieval. *4)* Extensive experiments show that the proposed method outperforms all SOTA methods on two benchmarks.

Related Work

Cross-modal Retrieval is a crucial task in vision-language domain (Yan et al. 2020a, 2021a,b, 2020b; Tu et al. 2023, 2024c,a; Zhang et al. 2024b), which aims to find relevant images by giving a caption. With advances in deep learning (Zha et al. 2019; Zhang et al. 2020a; Li et al. 2022; Liu et al. 2022; Tu et al. 2024b; Zhang et al. 2024a; Cui et al. 2024), researchers associate semantic relevant images and captions by projecting them into a joint embedding space to produce closer features. The visual feature has changed from coarse-grained one based on CNN (Faghri et al. 2018) to fine-grained one based on object detectors (Karpathy and Fei-Fei 2015; Lee et al. 2018; Li et al. 2019; Wang et al. 2020a; Zhang et al. 2020b; Kim, Kim, and Kwak 2023). Based on object detectors, some methods (Lee et al. 2018; Song and Soleymani 2019; Yan et al. 2022) apply attention mechanisms to focus on details. Other works (Li et al. 2019; Wang et al. 2020b) use GCN (Kipf and Welling 2016) to produce discriminative features by establishing relational graphs. However, the above object detector-based methods fail to detect scene text, resulting inefficiency in text-related scenarios. In contrast, HOPID regards scene texts in images as auxiliary clues for reducing the modality discrepancy.

Scene Text in Vision and Language provides fine-grained details for multi-modal tasks, such as Text-VQA (Biten et al. 2019; Yang et al. 2021; Liu et al. 2023) and Text-Caption (Sidorov et al. 2020; Yang et al. 2021). These tasks require model to comprehend scene text by connecting it with other visual elements. Recently, StacMR (Mafla et al. 2021) introduced scene-text aware cross-modal retrieval task with a corresponding dataset. To improve retrieval in both text-aware and text-free scenarios, ViSTA (Cheng et al. 2022) uses scene text as a third modality, fusing it with vision via a token and dual contrastive losses. Other works (Zhou et al. 2022, 2023) propose to align the scene text in images with captions locally, which improves the text-to-image retrieval performance and facility interpretability. More recently, researchers (Miyawaki et al. 2022) propose to build the relationship between scene text and captions through a co-mask strategy. In addition, VISTA (Zhou et al. 2024) can encode a variety of data types via an image tokenizer. However, these approaches do not explore the heterogeneous concept of scene text in cross-modal retrieval, therefore, we propose a framework to mine multi-modal heterogeneous scene text to narrow the cross-modal gap.

Methodology

Overall Model Architecture

The overall framework of our HOPID is illustrated in Figure 2. Given a raw image-caption pair, we first employ the visual and text prompt reasoning for pre-processing. Then we extract the visual and textual features via the OCR-aware

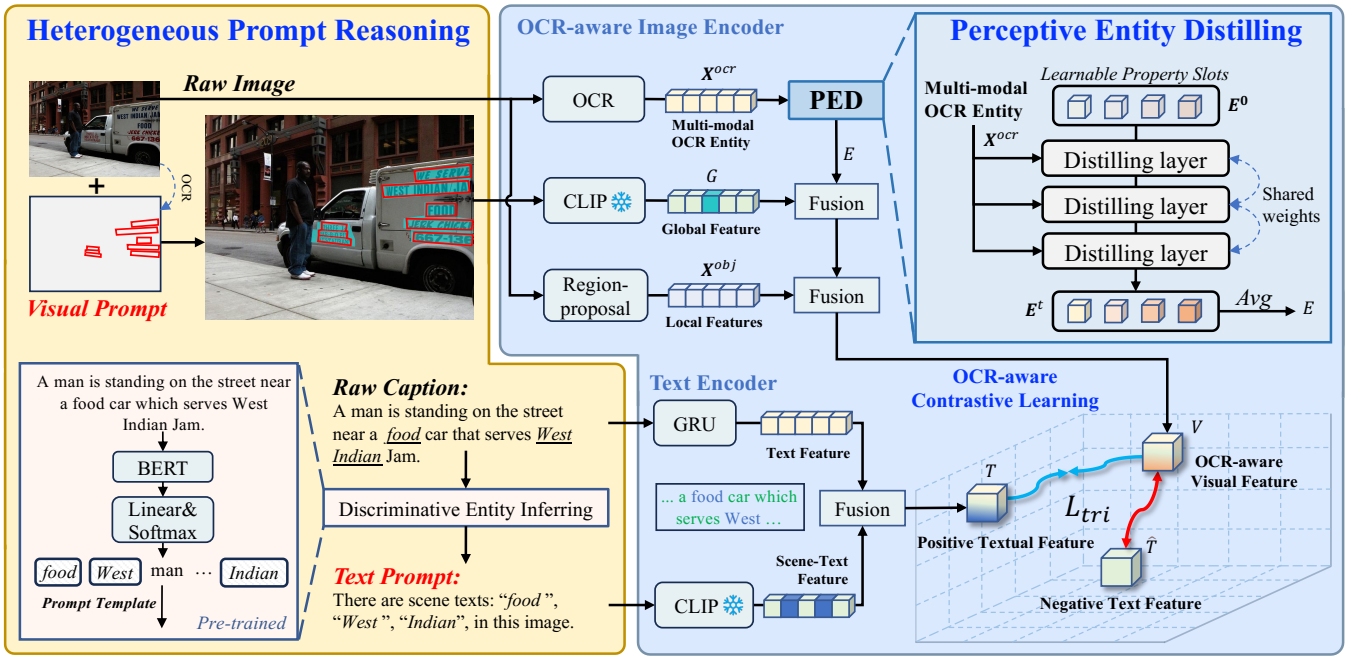

Figure 2: An overview of **HOPID** for StacMR task. First, the discriminative entity inferring module predicts scene text words to build text prompt. Meanwhile, visual prompt is generated based on OCR tools. Then OCR-aware image encoder and text encoder extract visual feature V and textual feature T. Finally, OCR-aware contrastive learning is used to align them.

image encoder and the text encoder, respectively. Finally, OCR-aware contrastive learning is employed to align the image-caption pairs in the common space.

Feature Extraction

Visual Representation. In OCR-aware image encoder, we first extract the global feature from the image pre-processed by visual prompt. And then we get local features and OCR features from the raw image. Finally, we fuse them to form the final visual representation V for a more comprehensive understanding of scene text from multiple perspectives.

Global features. Given an image pre-processed by the visual prompt, which hints at scene texts on pixel level, we utilize the frozen visual encoder of CLIP to extract the global feature. Thus, we obtain a 512D global feature G.

Local features. Given a raw image, we follow the conventional setting of the previous work (Mafla et al. 2021) to extract local features by Faster-RCNN (Ren et al. 2015). Thus we obtain the local features $\mathbf{X}^{obj} = \{x_m^{obj}\}_{m=1}^{M}$, where M is the maximum number of objects in the image.

OCR features. Given a raw image, we propose to extract OCR features from various modalities instead of semantic-only, we first employ an OCR system to obtain the words and bounding boxes as $\mathcal{S} = \{\boldsymbol{s}_n^{word}, \boldsymbol{s}_n^{bbox}\}_{n=1}^{N}$, where N is the maximum number of OCR tokens. According to $\mathcal{S}$, we represent each OCR token with following various modalities: (i) 300D word semantic embedding x_n^{ft} by FastText (Bojanowski et al. 2017), (ii) 2048D appearance feature x_n^{fr} of token's bounding box from Faster-RCNN (Ren et al. 2015), and (iii) 4D bounding box positional feature x_n^{bbox}. Then the final 2048D OCR features $\mathbf{X}^{ocr} = \{x_n^{ocr}\}_{n=1}^{N}$ are:

$$x_n^{ocr} = \sigma(LN(\boldsymbol{W}_1 \boldsymbol{x}_n^{ft} + \boldsymbol{W}_2 \boldsymbol{x}_n^{fr})) + \sigma((\boldsymbol{W}_3 \boldsymbol{x}_n^{bbox})), \tag{1}$$

where W_1, W_2 and W_3 are learned projection matrices, $LN(\cdot)$ is layer normalization and σ is a LeakyReLU activation function. After obtaining the unorganized OCR features $\mathbf{X}^{ocr}$ from various modalities ($i.e.$, semantics, visual and position), we feed them into the PED module discussed later, where OCR features compete with each other by slot attention, to learn a more robust scene text representation E.

For local features $\mathbf{X}^{obj}$, we construct a fully-connected graph as $G = \{\mathbf{X}^{obj}, R\}$, where R is the affinity matrix describing the relationship between each pair of local features:

$$R(x_i^{obj}, x_j^{obj}) = (W_a x_i^{obj})^T (W_b x_j^{obj}), \tag{2}$$

where W_a and W_b are learnable matrices. Then we use the graph convolutional network (GCN) to encapsulate high-level semantics about their relationships. In detail, the single layer of GCN is performed as below:

$$\mathbf{X}_l^{obj} = W_r(RX_{l-1}^{obj} W_g) + \mathbf{X}_{l-1}^{obj}, \tag{3}$$

where W_r and W_g are learnable matrices, l is the number of GCN layers, we set l to 4 here. Then the GRU is employed to consider the contextual information of objects and derive the high-level semantic local feature I_{obj} as:

$$I_{obj} = GRU(X_l^{obj}), \tag{4}$$

here *GRU* refers to the gated recurrent unit (Chung et al. 2014). Finally, we fuse the global feature G, the high-level

semantics local feature I_{obj} and the robust scene text feature E to obtain the final visual representation V as below:

$$V = (E \odot G + G) \odot I_{obj} + I_{obj}, \tag{5}$$

here, $\odot$ denotes the element-wise product. Therefore, we get the final visual representation V from multiple perspectives.

Text Representation. In the text encoder, we first use CLIP's text encoder to extract the discriminative scene-text feature from the text prompt discussed later. Then we get the text feature from the raw caption by gated recurrent unit (GRU). Finally, we fuse them by adding operation and obtain the final textual representation T.

Heterogeneous Prompt Reasoning

To mine the inherent connection of scene text in image and caption, we introduce heterogeneous prompt learning via visual and text prompt. Visual prompt provides context of scene text to alleviate local noise, while text prompt predicts discriminative scene text words in caption for alignment.

Visual Prompt. To enhance visual context of scene text in images which alleviates local noise, we devise a visual prompt to save its context at the pixel level. Given an image that contains scene texts, we detect their position as $\{\boldsymbol{s}_n^{bbox}\}_{n=1}^N = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}_{n=1}^N$ by the OCR tool (here we use paddleOCR). Then we generate the visual prompt by locating scene text in the image with a visual clue marker according to the position $\boldsymbol{s}_n^{bbox} \in \mathbb{R}^{4\times 2}$. This visual prompt hints at the positional information and the surroundings information like background for reasoning to alleviate local noise. In Table 4, we also provide an empirical analysis for different shapes of visual clue markers (*i.e.*, circle, square). Meanwhile, we introduce text prompt reasoning to reduce the cross-modal disparity of scene text by aligning them via heterogeneous prompt learning.

Text Prompt. To align the semantically identical scene-text that exists in both images and captions for narrowing the cross-modal gap, we first utilize the proposed DEI module (discussed below) to identify discriminative scene-text tokens in captions, a process that mimics human reasoning. Then, we generate the text prompt with a pre-defined prompt template for extracting scene-text features in caption.

Discriminative Entity Inferring. DEI aims to predict the discriminative scene-text tokens in captions as humans do for entities reasoning. For example, given the caption *"He wears jersey No.63."*, humans can easily detect *"63"* as the discriminative scene-text token. Technically, DEI is composed of a pre-trained BERT, followed by two MLP layers and a softmax layer. DEI is designed to project the input into two probability values enabling binary classification to discern whether the word in captions is an OCR token. Then we fine-tune DEI on the TextCaps-OCR (Zhou et al. 2022) dataset and utilize the cross-entropy of the softmax's output and the real label (1 for OCR and 0 for non-OCR) of words as the loss. Note that DEI is pre-trained and particularly used for predicting scene-text words in captions. Also, we conduct experiments to independently assess the DEI module's accuracy in predicting scene text tokens . After obtaining scene-text tokens $\mathbf{S}$ in captions, we integrate them with the pre-defined *Prompt Template* to build the text prompt which mimics human thinking: *"There are scene texts: [S] in this image."* Here, we follow the interesting finding in CLIP (Radford et al. 2021), putting quotes around the text to be recognized can improve performance in OCR task. When there is only one scene-text token, we use prompt template *"There is scene text: [S] in this image."* as a replacement. If there is no scene text detected from captions, we use the prompt template *"There may not be any scene text in this image."* instead. We explore more prompt designs in Table 5.

Perceptive Entity Distilling

Although multi-modal OCR entities $\mathbf{X}^{ocr} = \{x_n^{ocr}\}_{n=1}^N$ contain scene-text information from various modalities (*i.e.*, visual, semantic and position), they are unorganized and uncompact due to the correlation among them is not built, and it will cause the semantic ambiguity. To alleviate this problem, we introduce a Perceptive Entity Distilling (PED) module, which consists of multiple distilling layers that leverage the slot attention mechanism (Locatello et al. 2020). Slot attention is originally proposed in the object discovery task. Specifically, it initializes a set of learnable vectors as object slots, then object slots are used to interact with inputs and iteratively group information belonging to the same object. Inspired by its grouping characteristic, we propose the PED module to distill valuable scene-text information by perceiving multi-modal OCR context from multiple attributes and learn property-centric features as the final representation.

Specifically, we distill a set of L property-centric slots from OCR features $\mathbf{X}^{ocr}$ using a set of L learnable slots and the slot attention mechanism (Locatello et al. 2020). First, PED initializes L learnable property slots $\mathbf{E}^0 \in \mathbb{R}^{L\times D}$ from a Gaussian distribution $\mathcal{N}(0, 1)$, D is the embedding dimension of each slot. Then, multiple weights-shared distilling layers are introduced to iteratively distill the slots,

$$\mathbf{E}^t = \varphi(\mathbf{E}^{t-1}; \mathbf{X}^{ocr}) \in \mathbb{R}^{L\times D}, \tag{6}$$

where $\varphi(\cdot)$ is the cross-attention layer. In detail, we first norm $\mathbf{E}^{t-1}$ and $\mathbf{X}^{ocr}$ by layer normalization. Then, the input OCR features $\mathbf{X}^{ocr}$ are projected as $\mathbf{K} \in \mathbb{R}^{N\times D_h}$, $\mathbf{V} \in \mathbb{R}^{N\times D_h}$, and $\mathbf{E}^{t-1}$ are projected to $\mathbf{Q} \in \mathbb{R}^{L\times D_h}$. D_h is the dimension of the hidden state. Afterward, the attention between them is calculated as follows:

$$\mathbf{A}^t = \text{Softmax}\left(\frac{\mathbf{K}\mathbf{Q}^T}{\sqrt{D_h}}\right) \in \mathbb{R}^{N\times L}, \tag{7}$$

herein, Softmax$(\cdot)$ is calculated over property-slots direction. This way lets slots compete with each other, encouraging them to focus on distinct properties. With the t-th relationship matrix $\mathbf{A}^t$, PED updates slots $\mathbf{E}^t$ as follows:

$$\begin{aligned} \overline{\mathbf{E}}^t &= (\hat{\mathbf{A}}^t)^\top \mathbf{V}\mathbf{W}_o + \mathbf{E}^{t-1}, \text{where} \hat{\mathbf{A}}_{n,k}^t = \frac{\mathbf{A}_{n,k}^t}{\sum_{n=1}^N \mathbf{A}_{n,k}^t}, \\ \mathbf{E}^t &= \varphi(\mathbf{E}^{t-1}; \mathbf{X}^{ocr}) = \text{FFN}(\overline{\mathbf{E}}^t) + \overline{\mathbf{E}}^t, \end{aligned} \tag{8}$$

where $\mathbf{W}_o$ denotes a learnable projection matrix, and FFN refers to the feed-forward neural network. FFN is composed of a linear layer, GELU activation, and a normalization layer.

METHOD	CTC-1K							CTC-5K						
	IMG→TEXT			TEXT→IMG			RSUM	IMG→TEXT			TEXT→IMG			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
SCAN (Lee et al. 2018)	36.3	63.7	75.2	26.6	53.6	65.3	320.7	22.8	45.6	54.3	12.3	28.6	39.9	203.5
VSRN (Li et al. 2019)	38.2	67.4	79.1	26.6	54.2	66.2	331.7	23.7	47.6	59.1	14.9	34.7	45.5	225.5
STARNet (Mafla et al. 2021)	44.1	74.8	82.7	31.5	60.8	72.4	366.3	26.4	51.1	63.9	17.1	37.4	48.3	244.2
Dise-OCR (Zhou et al. 2022)	-	-	-	35.4	68	74.9	178.3	-	-	-	-	-	-	-
ViSTA-S (Cheng et al. 2022)	52.5	77.9	87.2	36.7	66.2	77.8	398.3	31.8	56.6	67.8	20	42.9	54.4	273.5
HM (Zhou et al. 2023)	-	-	-	38.2	69.4	79.7	187.3	-	-	-	-	-	-	-
VISTA (Zhou et al. 2024)	37.5	62.5	73.9	39.4	65.6	74.8	353.7	20.4	39.1	49.6	24.5	45.4	55.4	234.4
HOPID	**61.8**	**86.9**	**93.3**	**46.1**	**72.1**	**82.1**	**442.4**	**42.2**	**68.2**	**77.4**	**27.7**	**51.5**	**61.6**	**328.6**

Table 1: Comparisons with SOTA methods on CTC dataset.

METHOD	IMG→TEXT			TEXT→IMG			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
VSRN	14.3	34.9	46.2	9.5	26.2	37.2	168.3
SCAN	23.2	50.5	63.5	14.1	37.6	52.1	241.0
STARNet	28.7	53.7	65.1	19.8	40.1	51.6	259.0
Dual Encoder	52.9	76.4	83.2	40.5	63.0	71.1	387.1
VISTA	39.1	59.2	68.8	37.9	57.5	66.1	328.6
HOPID	**63.9**	**83.5**	**89.2**	**49.5**	**70.7**	**78.6**	**435.4**

Table 2: Comparisons with SOTA methods on TextCaps.

After these iterative updates, PED distills valuable entities among unorganized multi-modal OCR entities by perceiving multiple attributes at a fine-grained level to alleviate semantic ambiguity. Then, we use the average pooling to capture the overall representation of slots and obtain the robust scene text representation as $E = \text{Avg}(\mathbf{E}^t)$.

Training Objective

OCR-aware Contrastive Learning. To align the final OCR-aware visual feature V with the final textual feature T in a joint embedding space, we employ a triple ranking loss (Faghri et al. 2018; Lee et al. 2018), defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{tri}}(V,T) = &\sum_{\hat{T}} \Big[S(V,\hat{T}) - S(V,T) + \alpha\Big]_+ \\ &+ \sum_{\hat{V}} \Big[S(\hat{V},T) - S(V,T) + \alpha\Big]_+ ,\end{aligned} \tag{9}$$

where (V,T) is a positive image-text pair, $(V,\hat{T})$ and $(\hat{V},T)$ are negative image-text pairs in the batch, $\alpha > 0$ is a margin parameter, $[x]_+ = \max(x,0)$ and $S(\cdot)$ is the cosine similarity function. Given a positive pair (V,T), the hardest negatives are calculated by $\hat{V} = argmax_{z \neq V} S(z,T)$ and $\hat{T} = argmax_{d \neq T} S(V,d)$ in a mini-batch.

Experiments

Datasets

We conduct the experiments on two cross-modal retrieval datasets: COCO-Text Captioned (CTC) (Mafla et al. 2021) dataset and TextCaps (Sidorov et al. 2020) dataset. CTC contains two test sets, CTC-1K and CTC-5K. For fair comparisons, we strictly follow its previous split. On TextCaps, following before SOTA method (Miyawaki et al. 2022), we use 21,953 images for training and 3,166 images for testing.

Evaluation Metric

For evaluation, we use Recall@K, defined as the percentage of queries with correctly matched in the top-k retrieval results. We also report RSUM, which is the sum of Recall@K at $K \in \{1, 5, 10\}$ in both image-text and text-image retrieval.

Implementation Details

We set the number of iterations of PED t = 2, the dimension of each slot and OCR feature as $D = 2048$. Following Mafla et al. (2021), we set the maximum number of OCR tokens N = 20 and the maximum number of objects M = 36. In addition, we use paddleOCR to detect and recognize scene texts in images, selecting top-20 confident results. In cases where the number of OCR tokens is less than 20, we use zero-padding instead. Furthermore, the batch size is set to 300 and the model is trained and evaluated on one RTX 4090 GPU for 30 epochs. Adam optimizer is used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and a learning rate of $2e-4$.

Comparison with State-of-the-Art

In this section, we present the performance comparisons of HOPID and recent SOTA methods in Table 1 and Table 2.

Results on the CTC Dataset. As shown in Table 1, our HOPID achieves the best performance on both CTC-1K and CTC-5K test sets. We can observe that, compared with using scene text independently, utilizing both local and global alignment can better release the potential of heterogeneous scene text. Our HOPID leverages the pre-trained CLIP model which performs poorly in CTC dataset, while HOPID achieves a substantial performance improvement, reaching a margin of +58.5 in RSUM compared to the single CLIP model. Moreover, by distilling property-centric features from scene text through PED, HOPID outperforms those methods that do not explicitly filter scene text, it improves I2T-R@1 on CTC-1K by +9.3% compared to ViSTA.

Results on the TextCaps Dataset. In Table 2, HOPID outperforms all previous SOTA methods on TextCaps dataset, even VISTA, which uses EVA-CLIP with retraining with additional data. In detail, our HOPID brings substantial improvement across all metrics and outperforms comparison methods with RSUM = 435.4. This improvement benefits from heterogeneous prompt learning, which captures diverse manifestations of the scene text for alignment. Our HOPID still surpasses Dual Encoder (Miyawaki et al. 2022) which models OCR from single-modality, with a margin of +11%

	Ablation	CTC-1K							CTC-5K						
		IMG→TEXT			TEXT→IMG			RSUM	IMG→TEXT			TEXT→IMG			RSUM
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
A	Baseline	44.1	73.3	82.6	30.3	59.0	71.6	360.9	24.8	50.9	63.8	16.5	36.4	47.4	239.8
B	Baseline+PED	44.8	74.7	84.3	30.5	58.7	72.0	365.0	26.2	51.2	64.1	16.3	36.3	46.9	240.8
C	Baseline+T	48.3	75.8	86.2	31.2	60.8	72.5	374.8	27.4	53.2	65.0	15.9	35.1	45.9	242.5
D	Baseline+V	51.4	79.7	87.1	35.9	64.4	76.4	394.9	31.9	58.3	70.1	21.0	42.8	53.9	277.9
E	Baseline+O+T	52.1	77.0	86.9	37.2	65.8	77.1	396.1	33.9	60.1	70.4	20.4	41.9	52.8	279.6
F	Baseline+V+T	53.7	81.9	89.9	37.8	66.6	77.5	407.4	34.6	60.2	71.4	21.3	43.2	54.4	285.1
G	Baseline+O+T+PED	52.2	79.0	87.6	38.3	67.0	78.2	402.3	35.1	59.0	69.2	21.0	43.0	53.6	280.9
H	Baseline+V+T+PED	**61.8**	**86.9**	**93.3**	**46.1**	**72.1**	**82.1**	**442.3**	**42.2**	**68.2**	**77.4**	**27.7**	**51.5**	**61.6**	**328.6**

Table 3: Ablation study on both CTC-1K and CTC-5K. "(T) Text prompt", "(V) Visual prompt", "(O) Original image feature".

Shape	Quadrilateral	Circle	Cross	Square	Arrow
RSUM	439.4	**442.3**	437.1	439.7	439.7

Table 4: Comparisons with different visual prompt shapes.

	Text Prompt Template	I2T@5	T2I@5
1	without text prompt	78.5	64.9
2	There is no ST in this image.	78.4(-0.1)	62.7(-2.2)
3	This photo can detect ST: [**S**].	78.2(-0.3)	65.1(+0.2)
4	This photo contains ST: [**S**].	78.9(+0.4)	66.0(+1.1)
5	There are ST: [**S**] in this image.	**79.0(+0.5)**	**67.0(+2.1)**

Table 5: Comparisons with different text prompt templates on CTC-1K. ST is short for scene texts. **S** denote scene text words predicted by DEI from captions.

in I2T-R@1. This indicates that PED module can effectively perceive scene text from multiple perspectives (*i.e.*, visual, semantics, and position) to learn robust scene text features.

Ablation Study and Analysis

To further investigate the specific effects of each module in the proposed method, we conduct ablation studies on CTC-1K and CTC-5K in Table 3. Here, the "Baseline" refers to the model without the PED, the visual prompt, and the text prompt components. The "O" means the original global feature instead of the one with the visual prompt. After that, more analyses of our modules are presented.

Ablation Study of Text Prompt. In models 'A and C', when adding the text prompt, the RSUM is increased by +13.9% and +2.7% on CTC-1K and CTC-5K, respectively, which shows that the text prompt can obtain a crucial discriminative entity feature from captions. In models 'D and F', combining the text prompt with the visual prompt leads to an obvious gain. We ascribe this improvement to heterogeneous prompt learning, which aligns scene text from image and caption by leveraging the heterogeneous concepts.

Ablation Study of Visual Prompt. In models 'A and D', we find that the visual prompt contributes a large improvement in all metrics, indicating that it enriches visual context. Concretely, visual context mitigates the local noise by providing surrounding information about scene text. Compared with the original global feature (model 'E'), we note that the visual prompt (model 'F') elevates RSUM from 396.1 to 407.4 in CTC-1K. This validates that the visual prompt combined with the text prompt can hint the location and the scene text itself which was previously overlooked. In models 'G and H', we observe a notable increase across all metrics. This is attributed to the heterogeneous prompt learning and our PED, which reduces the cross-modal disparity of scene text.

Ablation Study of the PED Module. In models 'A and B' together with 'E and G', we find a little performance gain on the RSUM metric. This means that PED distills and integrates information from the unorganized and uncompact OCR features, thereby focusing on property-centric representation by perceiving multiple attributes at a fine-grained level to ease semantic ambiguity. In models 'F and H', combining PED with the heterogeneous prompts (model 'H') significantly outperforms the one without PED (model 'F'). This shows that PED is more effective at distilling valuable entities of scene text when working in conjunction with visual prompt. In detail, visual prompt enhances PED's ability to perceive entity attributes by providing visual context of scene text. To further investigate the synergy between PED and visual prompt, we visualize attention maps in Figure 3.

Effect of Different Templates of Text Prompt. In Table 5, we report the performance of five different prompt templates to explore the effect of text prompt on CTC-1K. The experiment setting follows 'G' in ablation study. First, we observe that various text prompt strategies (rows 3 to 5) outperform the baseline (row 1) without any text prompt, achieving improvements of +0.2%, +1.1%, and +2.1% in T2I@5, respectively. Indeed, our text prompt strategy consists of prompt template and scene text words **S** predicted by DEI from captions. To further clarify whether the improvement is made by discriminative scene text words **S** or the prompt template itself, we experiment with the toxic text prompt (row 2) template for all image-text pairs. The performance of the toxic prompt is worse than the one without text prompt (row 1) and decreases by 2.2% in T2I@5, indicating the toxic text prompt is harmful to our retrieval task. Therefore, it highlights the significance of scene text words **S** predicted by DEI module from captions, which are then used to construct text prompts with pre-defined prompt templates for cross-modal alignment. Moreover, our prompt template (row 5) exceeds the others (row 3) with a gain of +0.8% in I2T@5. We conjecture that the 'detect' in the prompt template (row 3) might mislead the model into treating it as a scene-text detection task, resulting in excessive focus on scene text rather than regarding it as auxiliary information for retrieval.

Figure 3: Visualization of attention maps of two examples.

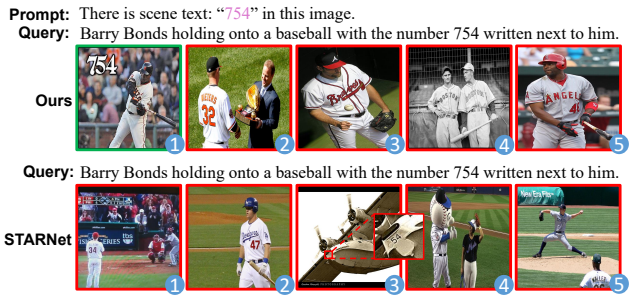

Figure 4: Top-5 text-to-image retrieval results.

Effect of Different Types of Visual Prompt. We conduct experiments with different shapes of visual clue markers for visual prompt on CTC-1K in Table 4, following experiment setting 'H' in ablation study. We observe that the red circle serves as a more effective shape for visual prompt. Thus we conjecture that the red circle mitigates local noise more efficiently by hinting at more accurate location information of scene text and reserving more surroundings information.

Effect of Slots Number in the PED Module. In this experiment, we investigate how the number of learnable property slots in the PED module influences the property-centric representation. The experiment setting is the same as 'G' in the ablation study. As depicted in Figure 5, the result shows that too many property slots ($N > 10$) lead to little distinction between similar property entities. This phenomenon makes it difficult to distill valuable scene text entities. Furthermore, it is noteworthy that when $N = 1$, the model struggles to perceive semantic ambiguity encountering difficulties in choosing discriminative property. Therefore, we set $N = 10$, a suitable number of property slots that alleviates semantic ambiguity arising from unorganized and uncompact OCR features by perceiving multiple attributes of the entity.

Evaluation of the DEI Module. To evaluate the accuracy of DEI in predicting OCR words from captions, we experiment on TextCaps-OCR. The results show that DEI reaches a precision of 88.6% for OCR words and 98.0% for non-OCR words, providing a reliable guarantee for text prompt.

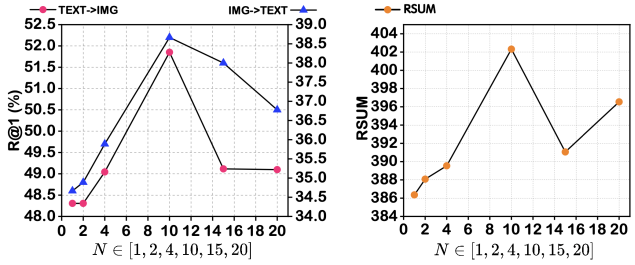

Figure 5: Comparisons with the number of property slots in PED evaluated on CTC-1K.

Qualitative Analysis

Attention Maps. We provide the visualization examples to discuss how the contextual information of scene text and property-centric representation benefit our task. In Figure 3 (a), we can see the model attends to shop signs ('King's' and 'Cross') and the subway sign ('underground'), corresponding to the 'sign' in text query. We argue that it benefits from the ability of PED to focus on properties and perceive similar attributes scene-text entities ('sign' in the image). In addition, with the assistance of the visual prompt, the region ('King's Cross', row 2 in (a)) is further emphasized based on visual context cues, showing the synergy between PED and visual prompt. In Figure 3 (b), we can see that the attention map without visual prompt overlooks the 'Police' in the image. However, with the inclusion of visual prompt, the position information of scene text is hinted. Through heterogeneous prompt learning, our model adjusts its attention to the scene text that was previously disregarded.

Retrieval Results. Figure 4 shows the top-5 retrieval results of text-to-image. The number on the bottom right of each image indicates ranking. Our HOPID retrieves the correct image in rank 1, while the STARNet (Mafla et al. 2021) cannot get it. In particular, the image ranked third by STARNet exhibits the semantic ambiguity issue. The number '54' here refers to the one on the airplane, but the query means the '754' written next to Barry Bonds. Indeed, this issue is often caused by semantic-only and unorganized OCR representation, while our PED module can alleviate this phenomenon by learning the property-centric representation of scene text in image for comprehensively understanding scene text.

Conclusion

This paper proposes a heterogeneous prompt-guided network to explore the heterogeneous scene text in images and captions and obtain property-centric scene text representation to address the challenges of understanding scene text in cross-modal retrieval. DEI infers discriminative entities from captions to build text prompt, benefiting retrieval, while visual prompt reduces local noise via marking scene texts in images. This heterogeneous prompt learning narrows the cross-modal gap of scene text. Moreover, PED distills salutary information to obtain a robust scene text representation. Our HOPID achieves state-of-the-art results on two scene-text aware cross-modal retrieval datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (U21B2024, 62322211, 62336008), "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (2023C01046, 2024C01023).

References

Biten, A. F.; Tito, R.; Mafla, A.; i Bigorda, L. G.; Rusiñol, M.; Jawahar, C. V.; Valveny, E.; and Karatzas, D. 2019. Scene Text Visual Question Answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 4290–4300. IEEE.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguistics*, 5: 135–146.

Cheng, M.; Sun, Y.; Wang, L.; Zhu, X.; Yao, K.; Chen, J.; Song, G.; Han, J.; Liu, J.; Ding, E.; and Wang, J. 2022. ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 5174–5183. IEEE.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR*, abs/1412.3555.

Cui, Y.; Li, L.; Zhang, J.; Yan, C.; Wang, H.; Wang, S.; Jin, H.; and Wu, L. 2024. Stochastic Context Consistency Reasoning for Domain Adaptive Object Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1331–1340.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, 12. BMVA Press.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 3128–3137. IEEE Computer Society.

Kim, D.; Kim, N.; and Kwak, S. 2023. Improving Cross-Modal Retrieval with Set of Diverse Embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 23422–23431. IEEE.

Kipf, T. N.; and Welling, M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR*, abs/1609.02907.

Koley, S.; Bhunia, A. K.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2024. You'll Never Walk Alone: A Sketch and Text Duet for Fine-Grained Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16509–16519.

Lee, K.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked Cross Attention for Image-Text Matching. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, 212–228. Springer.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual Semantic Reasoning for Image-Text Matching. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 4653–4661. IEEE.

Li, L.; Gao, X.; Deng, J.; Tu, Y.; Zha, Z.-J.; and Huang, Q. 2022. Long short-term relation transformer with global gating for video captioning. *IEEE Transactions on Image Processing*, 31: 2726–2738.

Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Li, Z.; Tian, Q.; and Huang, Q. 2022. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3003–3018.

Liu, Y.; Li, Z.; Yang, B.; Li, C.; Yin, X.; Liu, C.-l.; Jin, L.; and Bai, X. 2023. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; and Kipf, T. 2020. Object-Centric Learning with Slot Attention. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mafla, A.; de Rezende, R. S.; Gómez, L.; Larlus, D.; and Karatzas, D. 2021. StacMR: Scene-Text Aware Cross-Modal Retrieval. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, 2219–2229. IEEE.

Miyawaki, S.; Hasegawa, T.; Nishida, K.; Kato, T.; and Suzuki, J. 2022. Scene-Text Aware Image and Text Retrieval with Dual-Encoder. In Louvan, S.; Madotto, A.; and Madureira, B., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2022, Dublin, Ireland, May 22-27, 2022*, 422–433. Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 91–99.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In Vedaldi, A.; Bischof, H.; Brox, T.; and

Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, 742–758. Springer.

Song, Y.; and Soleymani, M. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 1979–1988. Computer Vision Foundation / IEEE.

Tu, Y.; Li, L.; Su, L.; Yan, C.; and Huang, Q. 2024a. Distractors-Immune Representation Learning with Cross-modal Contrastive Regularization for Change Captioning. In *ECCV*, 311–328.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024b. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2805–2815.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2024c. Context-aware Difference Distilling for Multi-change Captioning. *arXiv preprint arXiv:2405.20810*.

Wang, H.; Zhang, Y.; Ji, Z.; Pang, Y.; and Ma, L. 2020a. Consensus-Aware Visual-Semantic Embedding for Image-Text Matching. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV*, volume 12369 of *Lecture Notes in Computer Science*, 18–34. Springer.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2020b. Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, 1497–1506. IEEE.

Xie, Y.; Lin, Y.; Cai, W.; Xu, X.; Zhang, H.; Du, Y.; and He, S. 2024. D3still: Decoupled Differential Distillation for Asymmetric Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17181–17190.

Yan, C.; Gong, B.; Wei, Y.; and Gao, Y. 2020a. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4): 1445–1451.

Yan, C.; Hao, Y.; Li, L.; Yin, J.; Liu, A.; Mao, Z.; Chen, Z.; and Gao, X. 2021a. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video technology*, 32(1): 43–51.

Yan, C.; Li, Z.; Zhang, Y.; Liu, Y.; Ji, X.; and Zhang, Y. 2020b. Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(4): 1–17.

Yan, C.; Meng, L.; Li, L.; Zhang, J.; Wang, Z.; Yin, J.; Zhang, J.; Sun, Y.; and Zheng, B. 2022. Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s): 1–18.

Yan, C.; Teng, T.; Liu, Y.; Zhang, Y.; Wang, H.; and Ji, X. 2021b. Precise no-reference image quality evaluation based on distortion identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s): 1–21.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florêncio, D.; Wang, L.; Zhang, C.; Zhang, L.; and Luo, J. 2021. TAP: Text-Aware Pre-Training for Text-VQA and Text-Caption. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 8751–8761. Computer Vision Foundation / IEEE.

Zha, Z.-J.; Liu, D.; Zhang, H.; Zhang, Y.; and Wu, F. 2019. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 710–722.

Zhang, B.; Li, L.; Wang, S.; Cai, S.; Zha, Z.-J.; Tian, Q.; and Huang, Q. 2024a. Inductive state-relabeling adversarial active learning with heuristic clique rescaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020a. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020b. Context-Aware Attention Network for Image-Text Retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 3533–3542. Computer Vision Foundation / IEEE.

Zhang, Z.; Li, L.; Cong, G.; Yin, H.; Gao, Y.; Yan, C.; Hengel, A. v. d.; and Qi, Y. 2024b. From speaker to dubber: movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7523–7532.

Zhou, J.; Liu, Z.; Xiao, S.; Zhao, B.; and Xiong, Y. 2024. VISTA: Visualized Text Embedding For Universal Multi-Modal Retrieval. *CoRR*, abs/2406.04292.

Zhou, X.; Li, S.; Chen, H.; and Zhu, A. 2022. Disentangled OCR: A More Granular Information for "Text"-to-Image Retrieval. In Yu, S.; Zhang, Z.; Yuen, P. C.; Han, J.; Tan, T.; Guo, Y.; Lai, J.; and Zhang, J., eds., *Pattern Recognition and Computer Vision - 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4-7, 2022, Proceedings, Part I*, volume 13534 of *Lecture Notes in Computer Science*, 510–523. Springer.

Zhou, X.; Zhu, A.; Chen, H.; and Pan, W. 2023. Scene Text Involved "Text"-to-Image Retrieval through Logically Hierarchical Matching. In *IEEE International Conference on Multimedia and Expo, ICME 2023, Brisbane, Australia, July 10-14, 2023*, 114–119. IEEE.