MS-DPPs: Multi-Source Determinantal Point Processes for Contextual Diversity Refinement of Composite Attributes in Text to Image Retrieval

Naoya Sogi[1] , **Takashi Shibata**[1] , **Makoto Terao**[1] , **Masanori Suganuma**[2] , **Takayuki Okatani**[2,3]
[1]Visual Intelligence Research Laboratories, NEC Corporation
[2]Graduate School of Information Sciences, Tohoku University
[3]RIKEN Center for AIP
naoya-sogi@nec.com, t.shibata@ieee.org, m-terao@nec.com,
{suganuma, okatani}@vision.is.tohoku.ac.jp

Abstract

Result diversification (RD) is a crucial technique in Text-to-Image Retrieval for enhancing the efficiency of a practical application. Conventional methods focus solely on increasing the diversity metric of image appearances. However, the diversity metric and its desired value vary depending on the application, which limits the applications of RD. This paper proposes a novel task called CDR-CA (Contextual Diversity Refinement of Composite Attributes). CDR-CA aims to refine the diversities of multiple attributes, according to the application's context. To address this task, we propose Multi-Source DPPs, a simple yet strong baseline that extends the Determinantal Point Process (DPP) to multi-sources. We model MS-DPP as a single DPP model with a unified similarity matrix based on a manifold representation. We also introduce Tangent Normalization to reflect contexts. Extensive experiments demonstrate the effectiveness of the proposed method.

1 Introduction

Text-to-Image Retrieval (T2IR) aims to retrieve images related to input text from a set of images [Frome *et al.*, 2013; Faghri *et al.*, 2017; Lee *et al.*, 2018; Cao *et al.*, 2022; Wei *et al.*, 2024]. Recent advances in visual language models have become a new paradigm in the field of computer vision tasks, including T2IR [Chen *et al.*, 2020; Radford *et al.*, 2021; Li *et al.*, 2021; Li *et al.*, 2022; Wang *et al.*, 2023; Zeng *et al.*, 2024]. For example, BLIP-2 and its variants can perform searches with significantly higher accuracy than conventional methods [Li *et al.*, 2023; Sogi *et al.*, 2024a].

Result diversification (RD) is a fundamental task that aims to output relevant but diverse items so that users can comprehensively grasp the information within a gallery dataset [Zhang *et al.*, 2005; Ionescu *et al.*, 2016; Ionescu *et al.*, 2021]. RD methods are demanded for various applications [Affandi *et al.*, 2014; Ionescu *et al.*, 2021; Wu *et al.*, 2024] to enhance the efficiency of a practical T2IR application. The determinantal point process (DPP) is one of the standard approaches that retrieves the diverse samples based on the relevance score (i.e., similarity between the text query and images) and the appearance similarity between images [Kulesza and Taskar, 2012]. Conventional RD methods focus solely on increasing the diversity of image appearances.

Images typically contain various composite attributes such as shooting time and region, and the target attributes for RD should differ for each application. Furthermore, whether the diversity of each attribute is increased or decreased depends on the application. For example, if the goal is to retrieve images related to disaster damage and grasp the overall damages from the retrieved images [Weber *et al.*, 2022; Sogi *et al.*, 2024b], it is necessary to ensure high diversity in terms of both time and area. Conversely, if the goal is to analyze disaster damages in a local area deeply, it is effective to search for images taken in a concentrated location with diverse appearances. There is a strong demand for a versatile diversity-aware framework that can incorporate application-adaptive contexts in a plug-and-play manner while inheriting the strong capability of visual language models.

This paper proposes a generalized task called CDR-CA (Contextual Diversity Refinement of Composite Attributes), which aims to simultaneously refine the diversities of multiple attributes, such as image appearance and shooting time. To solve this task, we propose MS-DPPs (Multi-Source DPPs), a simple yet strong baseline that extends DPP to multiple sources. The key to MS-DPPs is that they can consider multiple attributes of each image in an application-adaptive and plug-and-play manner. We present that MS-DPP can be regarded as a DPP model with a unified similarity matrix, which is an interpolation of attributes' similarity matrices on the tangent space of the Symmetric Positive Definite (SPD) manifold. This formulation enables us to employ established optimization techniques from the standard DPPs. Consequently, MS-DPPs can refine the diversity of multiple attributes by reflecting each context while inheriting the strengths of existing retrieval frameworks, as shown in Fig. 1. For instance, in Fig. 1, CDR-CA requires us to provide a list of images with diverse appearance and shooting time.

In addition to retrieval accuracy and diversity measures, it is important to faithfully reflect attribute weights (or user preferences) in a user-interactive system. For instance, if a user increases the importance of an attribute, the system

[0]Code: https://github.com/NEC-N-SOGI/msdpp

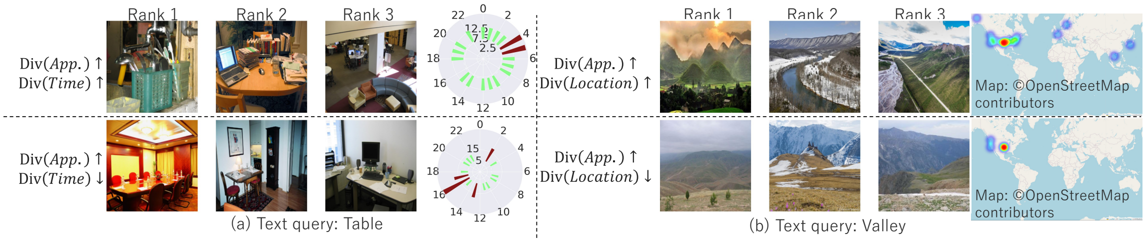

Figure 1: Examples by MS-DPP for four CDR-CA tasks. The left figures show shooting time refinement results, including the top three images and polar histograms of shooting times in the top 20 images. The upward bar of each histogram is at 0:00, and time moves clockwise, rounding the clock in 24 hours. The right figures show location refinement results, including heatmaps of the locations in the top 20 images.

should further improve its diversity. To address this issue, we introduce Tangent Normalization (TN) to MS-DPPs. TN is a normalization process before computing an interpolation of attributes' similarity matrices. This process suppresses the effect of norms of the similarity matrices in the tangent space so that attribute weights can be effectively reflected.

Extensive experiments demonstrate that our method outperforms the state-of-the-art methods in terms of mean of retrieval accuracy and diversity.

The contributions of this paper are fourfold: 1) We propose a novel problem setting called the CDR-CA task, which aims to refine the diversity of multiple attributes simultaneously. 2) We propose MS-DPPs, a simple yet strong baseline that extends the DPP manner to multiple sources. 3) We propose Tangent Normalization (TN) to reflect application contexts in MS-DPPs. 4) We confirmed the effectiveness of the proposed method using three publicly available datasets.

2 Related Works

2.1 Text to Image Retrieval

Text-to-image retrieval (T2IR) is a pivotal task in the field of vision and language research [Frome *et al.*, 2013; Faghri *et al.*, 2017; Lee *et al.*, 2018; Kaur *et al.*, 2021; Cao *et al.*, 2022]. T2IR aims to retrieve images from a vast database so that the retrieved images are relevant to a given text query. This task is typically achieved by predicting a relevance score for a pair of each image and the query text. Two primary approaches dominate relevance prediction: Image-Text Contrastive (ITC) [Radford *et al.*, 2021; Yu *et al.*, 2022; Chen *et al.*, 2024] and Image-Text Matching (ITM) methods [Faghri *et al.*, 2017; Lee *et al.*, 2018].

ITC determines relevance using a similarity measure between a query feature and an image feature, commonly using cosine similarity. This approach is recognized as fast and effective for T2IR. In contrast, ITM treats relevance prediction as a classification problem, assessing whether an image and a query form a matching pair. While ITM tends to be more accurate than ITC, ITM is also more computationally intensive. A hybrid approach is often employed [Li *et al.*, 2023] to leverage the advantages of the methods: ITC first retrieves the top-K images, which are then re-ranked by ITM.

Our MS-DPP refines the diversity in retrieval results by re-ranking the original results obtained by a T2IR method. As the re-ranking is a post-processing step, our method can be combined with any T2IR method.

2.2 Result Diversification

Result Diversification (RD) is a fundamental task in information retrieval [Carbonell and Goldstein, 1998; Zhang *et al.*, 2005; Vieira *et al.*, 2011; Ionescu *et al.*, 2016; Chen *et al.*, 2018; Ionescu *et al.*, 2021; Wu *et al.*, 2024]. This task aims to output relevant but diverse items so that users can comprehensively understand the information in a dataset.

Result Diversification in Image Retrieval

In the context of image retrieval, RD is a task that increases the diversity of image appearance in the top-K images while maintaining highly relevant images to a text query [van Leuken *et al.*, 2009; Boteanu *et al.*, 2015; Ionescu *et al.*, 2016; Bouhlel *et al.*, 2017; Boteanu *et al.*, 2017; Benavent *et al.*, 2019; Ionescu *et al.*, 2021; Tekli, 2022; Zhao *et al.*, 2023]. This task is important to provide a comprehensive view of the relevant images in a gallery database. There are three primary approaches to RD: 1) Maximal Marginal Relevance (MMR)-based methods [Carbonell and Goldstein, 1998], 2) Clustering-based methods [Boteanu *et al.*, 2017], and 3) Determinantal Point Process (DPP)-based methods [Odile, 1975; Kulesza and Taskar, 2011]. These methods are post-processing techniques, i.e., they re-rank images using relevance scores and image features.

MMR-based methods sequentially select an image that maximizes the sum of relevance to a query and a diversity measure between a candidate and previously selected images. This method is widely used in RD, as it is simple and effective. Clustering-based methods sequentially select an image from each cluster while referring to relevances. This approach significantly increases diversity, although it decreases retrieval accuracy. To improve the trade-off, some extensions are proposed, e.g., using pseudo-labeling. DPP-based methods use a probabilistic model on subsets of all images. A probabilistic model is defined by a matrix determinant and assigns a high probability to a relevant and diverse subset. As DPPs are effective, they are widely used in RD and actively studied, such as to improve scalability [Chen *et al.*, 2018] and further understanding of their mechanism [Kulesza and Taskar, 2012; Mariet *et al.*, 2019].

We propose a novel task, Contextual Diversity Refinement of Composite Attributes (CDR-CA), to open up novel applications of RD in T2IR. CDR-CA is an extension of RD that refines the diversity of multiple attributes and has mixed diversity directions, i.e., increasing some attributes and decreasing others. We then propose Multi-Source DPP (MS-DPP) tailored method of DPPs to CDR-CA.

3 Preliminary

3.1 Result Diversification

Let us begin with an introduction to the standard result diversification task in image retrieval. Let $\{\mathbf{f}_i \in \mathbb{R}^{d_I}\}_{i=1}^{N_I}$ be image features of gallery images $\{I_i\}_{i=1}^{N_I}$ and q be a query text. Text-to-image retrieval methods output a top-K ranked list $\mathcal{Y}_K$ of images by calculating a relevance score r_i between the query text and each image I_i. The result diversification task is to increase the diversity of image appearance in the top-K ranked list, while maintaining high-relevance images. This concept can be written as follows:

$$\mathrm{argmax}_{\mathcal{Y}_K}(1-\theta)\sum_{i\in\mathcal{Y}_K} r_i + \theta\mathrm{Div}(\mathcal{Y}_K), \tag{1}$$

where θ is a trade-off parameter between retrieval accuracy and diversity, and $\mathrm{Div}(\mathcal{Y}_K)$ is a diversity measurement, such as the negative maximum pairwise similarity of images in $\mathcal{Y}_K$ [Wu *et al.*, 2024].

3.2 Result Diversification by DPPs

In the following, we describe a DPP-based method [Kulesza and Taskar, 2011], which is the basis of our method.

Notations

We introduce notations for DPPs. Let N_I be the number of images in a gallery, $\{\mathbf{f}_i \in \mathbb{R}^{d_I}\}_{i=1}^{N_I}$ be image features. A retrieval model outputs relevance scores $\{r_i\}_{i=1}^{N_I}$ between a query text and images. A retrieval result, i.e., top-K ranked list, is represented by an ordered index set $\mathcal{Y}_K$.

DPPs use three matrices: $\mathbf{R}$, $\mathbf{S}$, and $\mathbf{L}$. $\mathbf{R} \in \mathbb{R}^{N_I\times N_I} = \mathrm{diag}(r_1, r_2, ..., r_{N_I})$. $\mathbf{S} \in \mathbb{R}^{N_I\times N_I}$ is a similarity matrix of image features. $\mathbf{L} \in \mathbb{R}^{N_I\times N_I} = \mathbf{R}\,\mathbf{S}\,\mathbf{R}$ defines a DPP. Let $\mathbf{E}$ be an identity matrix. A matrix with subscript $\mathcal{Y}_g$ is a submatrix indexed by $\mathcal{Y}_g$. For instance, $\mathbf{S}_{\mathcal{Y}_g} \in \mathbb{R}^{g\times g}$ is a submatrix of $\mathbf{S}$ indexed by $\mathcal{Y}_g$: $(\mathbf{S}_{\mathcal{Y}_g})_{(i,j)} = (\mathbf{S})_{x,y}, x = \mathcal{Y}_g(i), y = \mathcal{Y}_g(j)$, where $\mathcal{Y}_g(i)$ is the ith element of $\mathcal{Y}_g$.

Let $\log$ be the natural logarithm, $\mathbf{logm}$ and $\mathbf{expm}$ be the matrix logarithm and matrix exponential, and $|\cdot|$ and $\mathrm{tr}\cdot$ be the matrix determinant and trace, respectively.

Determinantal Point Process (DPP)

DPP increases the diversity of image appearance in the top-K ranked list, while maintaining highly relevant images by using the following probabilistic model:

$$p(\mathcal{Y}_g) = \frac{|\mathbf{L}_{\mathcal{Y}_g}|}{|\mathbf{L}+\mathbf{E}|} \propto |\mathbf{L}_{\mathcal{Y}_g}| = |\mathbf{R}_{\mathcal{Y}_g}|\,|\mathbf{S}_{\mathcal{Y}_g}|\,|\mathbf{R}_{\mathcal{Y}_g}|. \tag{2}$$

Large determinants of $\mathbf{R}_{\mathcal{Y}_g}$ and $\mathbf{S}_{\mathcal{Y}_g}$, i.e., large probability, indicate that the images in $\mathcal{Y}_g$ are relevant and diverse, respectively. Please refer to the paper [Kulesza and Taskar, 2012] for the detailed mechanism. We can obtain a diversified image set as $\mathcal{Y}_g^* = \mathrm{argmax}_{\mathcal{Y}_g}|\mathbf{R}_{\mathcal{Y}_g}\mathbf{S}_{\mathcal{Y}_g}\mathbf{R}_{\mathcal{Y}_g}|$. This optimization problem is generally NP-hard, but we can obtain $\mathcal{Y}_g^*$ efficiently by sequentially selecting images that maximize the determinant [Kulesza and Taskar, 2011; Chen *et al.*, 2018].

	N_A	Diversity Direction s	Preference w
RD	1	$s=+1$	$w=1$
CDR-CA	>1	$s_i \in \{-1,+1\}$	$w_i \in [0,1]$

Table 1: Differences between the Result Diversification (RD) and the proposed CDR-CA tasks.

4 Problem Formulation: Contextual Diversity Refinement of Composite Attributes

Our Contextual Diversity Refinement of Composite Attributes (CDR-CA) task is to refine the diversity of multiple attributes, such as image appearance and shooting time, while reflecting relevance scores, as shown in Fig. 1. For example, CDR-CA requires us to provide a list of images that 1) have diverse appearances and shooting times or 2) have diverse appearances while taken at a concentrated time. The concept of CDR-CA can be written as follows:

$$\mathrm{argmax}_{\mathcal{Y}_K}(1-\theta)\sum_{i\in\mathcal{Y}_K} r_i + \theta\sum_{j=1}^{N_A} s_j w_j \mathrm{Div}_j(\mathcal{Y}_K), \tag{3}$$

where N_A is the number of attributes, $s_j \in \{-1,+1\}$ is an indicator parameter to determine diversity of jth attribute is increased (+1) or decreased (-1), w_j is the weight of jth attribute, and $\mathrm{Div}_j(\mathcal{Y}_K)$ is the diversity measurement of jth attribute of images in $\mathcal{Y}_K$. For simplicity, we add image appearance as the first attribute. Table 1 summarizes the differences between CDR-CA and the standard result diversification.

5 Proposed Method

This section presents the proposed method, Multi-Source Determinantal Point Process (MS-DPP), a novel extension of the traditional Determinantal Point Processes (DPPs) tailored to address CDR-CA. Figure 2 illustrates the overview of the proposed method. We first describe the basic formulation of MS-DPP using multiple DPPs to handle multiple attributes. We then demonstrate that the basic formulation can be seamlessly transformed into a single DPP model by unifying similarity matrices of multiple attributes through operations on the SPD manifold. This unification enables us to leverage established optimization techniques for DPPs. Finally, we introduce Tangent Normalization (TN), a novel technique that aligns attribute diversities with user-defined weights.

5.1 Multi-Source Determinantal Point Processes

Basic Idea

MS-DPP is an extension of the DPP framework to address a CDR-CA task. The key concept involves applying multiple DPPs across different attributes to enhance their diversity simultaneously. Let us first consider a simple CDR-CA task; all attributes hold equal importance ($w_i = 1$), and the diversity of each attribute is increased ($s_i = 1$). The basic model is defined as follows:

$$\prod_{i=1}^{N_A} p_i(\mathcal{Y}_g) \propto f_{base}(\mathcal{Y}_g) = \prod_{i=1}^{N_A} |\mathbf{R}_{\mathcal{Y}_g}|\,|\mathbf{S}_{i,\mathcal{Y}_g}|\,|\mathbf{R}_{\mathcal{Y}_g}|, \tag{4}$$

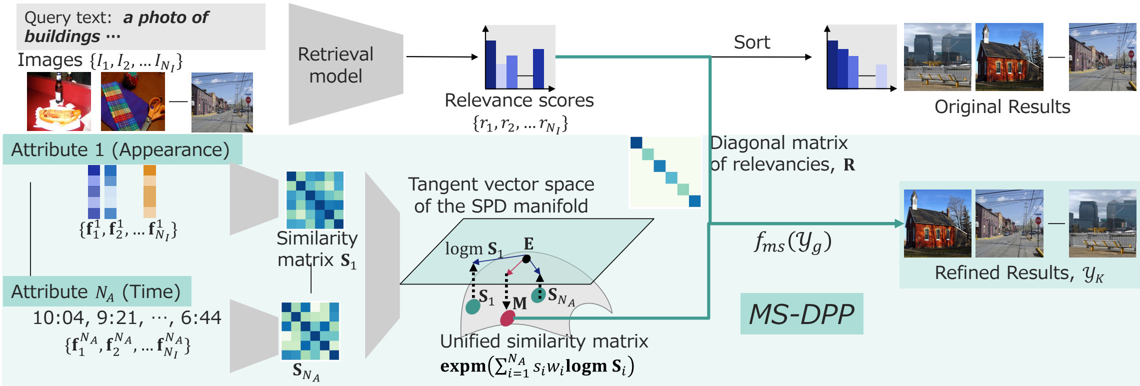

Figure 2: Overview of MS-DPP. We first calculate the similarity matrices of each attribute and then unify them into a single similarity matrix through operations, matrix logarithm and exponential, on the SPD manifold. Given the unified similarity matrix and the diagonal matrix of relevances, MS-DPP re-ranks images to refine the diversity of multiple attributes.

where p_i is a DPP model for the ith attribute defined in (2), $\mathbf{S}_i$ and $\mathbf{R}$ are a similarity matrix of the ith attribute and a diagonal matrix of relevances, respectively. Maximization of this composite model f_{base} refines the diversity of multiple attributes within the top-K ranked list.

To efficiently derive the optimal list $\mathcal{Y}_K^*$, we reformulate the composite DPP model f_{base} into a single DPP model by unifying the similarity matrices $\{\mathbf{S}_i\}$ into one SPD matrix $\mathbf{M}$. This unification allows us to employ optimization techniques [Kulesza and Taskar, 2011; Chen *et al.*, 2018] from the standard DPP framework.

Unified Representation for MS-DPP

In the following, we first reformulate the composite DPP model into a single DPP model with a unified similarity matrix. We then generalize the unified representation to accommodate various cases of CDR-CA tasks, i.e., mixed diversity directions $s_i \in \{-1, +1\}$ and flexible weights $w_i \in \mathbb{R}$.

Theorem 1. Unified Representation: *The composite DPP model (f_{base}) in (4) can be expressed as a single DPP model with a unified similarity matrix* $\mathbf{M}$ *as follows:*

$$f_{base}(\mathcal{Y}_g) \propto f_{unify}(\mathcal{Y}_g) = \left|\mathbf{R}_{\mathcal{Y}_g}\right| |\mathbf{M}| \left|\mathbf{R}_{\mathcal{Y}_g}\right|, \tag{5}$$

where $\mathbf{M} = \mathbf{expm} \sum_{i=1}^{N_A} \mathbf{logm}\, \mathbf{S}_{i,\mathcal{Y}_g}$.

Proof: Please refer to Sec. 2 in the supplementary material.

The essence of the unified representation lies in transforming the product of multiple matrix determinants into a single determinant by utilizing the matrix logarithm and exponential. Consequently, we can leverage existing optimization techniques from the standard DPP framework to efficiently derive the optimal result $\mathcal{Y}_K^*$ of f_{unify}.

Definition 1 (MS-DPP). In the above, we consider a simple CDR-CA task where all attributes have equal importance $w_i = 1$ and the diversity of each attribute is increased $s_i = 1$. To tackle various types of CDR-CA tasks, we extend the unified representation as follows:

$$f_{ms}(\mathcal{Y}_g) = \left|e^{\alpha\, \mathbf{R}_{\mathcal{Y}_g}}\right| \left|\mathbf{expm} \sum_{i=1}^{N_A} s_i w_i\, \mathbf{logm}\, \mathbf{S}_{i,\mathcal{Y}_g}\right| \left|e^{\alpha\, \mathbf{R}_{\mathcal{Y}_g}}\right|, \tag{6}$$

Algorithm 1 MS-DPP for a CDR-CA task

Input: Similarity matrices $\{\mathbf{S}_i \in \mathbb{R}^{N_I \times N_I}\}_{i=1}^{N_A}$, Relevance matrix $\mathbf{R} \in \mathbb{R}^{N_I \times N_I}$, attribute weights $\{w_i\}$, refinement directions $\{s_i\}$, and trade-off parameter θ.
Output: A top-K ranked list $\mathcal{Y}_K$.
Initialize $\mathcal{Y}_g = \emptyset, \mathcal{I} = \{1, 2, \dots, N_I\}$
while $|\mathcal{Y}_g| < K$ **do**
  for $j = 1$ **to** N_A **do**
    Calculate $f_j = f_{ms}(\mathcal{Y}_{g \cup j})$ (6).
  end for
  Update $\mathcal{Y}_g = \mathcal{Y}_g \cup \{j^*\}$, $\mathcal{I} = \mathcal{I} \setminus \{j^*\}$, where $j^* = \mathrm{argmax}_j f_j$.
end while

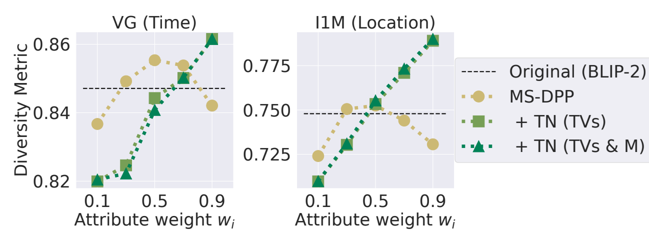

Figure 3: Transition of the diversity of an attribute varying its weight w_i from 0.1 to 0.9 with a step size of 0.2 with $\theta = 0.9$. The task is to increase diversity in appearance and an attribute (time or location). The higher diversity metric indicates better performance.

where $s_i \in \{-1, +1\}$ indicates the diversity refinement direction of the ith attribute, w_i is the weight or user's preference of the ith attribute, and $e^{\alpha\, \mathbf{R}_{\mathcal{Y}_g}}$ is elementwise exponential of $\alpha\, \mathbf{R}_{\mathcal{Y}_g}$. $\alpha = \frac{\theta}{2(1-\theta)}, \theta \in (0, 1)$ is a trade-off parameter [Chen *et al.*, 2018] between retrieval accuracy and diversity refinement. We refer to f_{ms} as an **MS-DPP** model. Algorithm 1 shows the k-DPP [Kulesza and Taskar, 2011] based optimization of an MS-DPP model. In the experiments, we use the computationally efficient extension [Chen *et al.*, 2018] of this algorithm.

5.2 MS-DPP with Tangent Normalization

Motivation and Basic Idea

Figure 3 shows the transition of the diversity of an attribute varying its weight w_i. The details of this experiment will be described in Sec. 6.1. MS-DPP improves the diversity metric, although MS-DPP with a large weight $w_i > 0.5$ does not always increase the metric as much as the weight. It is important for a user-interactive system to reflect weights faithfully. To address this issue, we introduce a novel normalization, Tangent Normalization (TN).

The reason for the above issue can be understood from the fact that the unification ($\sum_{i=1}^{N_A} s_i w_i \,\mathbf{logm}\, \mathbf{S}_{i,\mathcal{Y}_g}$) of similarity matrices can be regarded as an interpolation of the tangent vectors $\{\mathbf{logm}\, \mathbf{S}_{i,\mathcal{Y}_g}\}$ (TVs) on the SPD manifold. This interpretation arises because the matrix logarithm $\mathbf{logm}$ maps an SPD matrix to the tangent space, an Euclidean space, with the identity matrix as the origin (please refer to Sec. 1 in the supplementary material). This interpretation suggests that the norm of each TV can influence the balance of diversity refinement among attributes, i.e., an attribute with a large norm can be dominant, although a user makes large importance on other attributes.

Tangent Normalization

To mitigate the influence of the norm of each TV, we normalize TVs before taking the weighted sum. Considering that $\mathbf{R}$ is also an SPD matrix and we need to consider the trade-off between retrieval accuracy and diversity refinement, we propose the following normalization:

$$\mathbf{A}'_{i,\mathcal{Y}_g} = \frac{b}{\|\,\mathbf{logm}\, \mathbf{S}_{i,\mathcal{Y}_g}\|_F} \mathbf{logm}\, \mathbf{S}_{i,\mathcal{Y}_g}, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm, i.e., the norm in the tangent vector space, and $b = \|\,\mathbf{logm}\, \mathbf{R}_{\mathcal{Y}_g}\|_F$ is the norm of the tangent vector of the relevance matrix. We call this normalization as **Tangent Normalization (TN)**. Figure 4 shows the conceptual diagram of TN.

Additionally, we normalize the mean of normalized vectors ($\mathbf{M}' = \mathbf{expm} \sum_{i=1}^{N} s_i w_i \,\mathbf{A}'_{i,\mathcal{Y}_g}$) to faithfully reflect the accuracy-diversity trade-off, as follows:

$$\mathbf{logm}\, \mathbf{M}'' = \frac{b}{\|\,\mathbf{logm}\, \mathbf{M}'\|_F} \mathbf{logm}\, \mathbf{M}'. \tag{8}$$

We use $\mathbf{M}'$ or $\mathbf{M}''$ in the MS-DPP model (6) instead of $\mathbf{M}$ to reflect the user's preferences $\{w_i\}$ among attributes. As shown in Fig. 3, MS-DPP with TN improves the diversity metric as the weight increases. This improvement will be quantitatively confirmed in the experiment section using the evaluation index defined in the following subsection.

5.3 Preference Reflection Score (PRS)

To assess how well a user's preference w_i for an attribute is reflected, we introduce the Preference Reflection Score (PRS). Given refinement results $\{\mathcal{Y}_{g,j}\}_{j=1}^{T}$ and the user's preference $\{w_{i,j}\}_{j=1}^{T}$, ($w_{i,1} \le w_{i,2} \le \ldots \le w_{i,T}$), for the ith attribute in the jth result, we calculate the PRS as follows:

$$\mathrm{PRS} = \sum_{j=2}^{T} \frac{(\overline{\mathrm{Div}}_i(\mathcal{Y}_{g,j}) - \overline{\mathrm{Div}}_i(\mathcal{Y}_{g,j-1}))}{(w_{i,j} - w_{i,j-1})}, \tag{9}$$

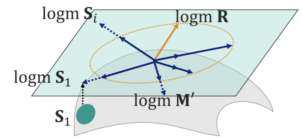

Figure 4: Conceptual diagram of Tangent Normalization (TN). To suppress the influence of the norm of each tangent vector, TN normalizes tangent vectors of similarity matrices $\{\mathbf{S}_i\}$ and the unified similarity matrix $\mathbf{M}'$ by the norm of the tangent vector of $\mathbf{R}$.

where $\overline{\mathrm{Div}}_i(\mathcal{Y}_{g,j})$ is the diversity measurement of the ith attribute in the top-g ranked list $\mathcal{Y}_{g,j}$. $\overline{\mathrm{Div}}_i$ is normalized to $\max_j \overline{\mathrm{Div}}_i(\mathcal{Y}_{g,j}) = 1, \min_j \overline{\mathrm{Div}}_i(\mathcal{Y}_{g,j}) = 0$. PRS increases when the diversity of the ith attribute grows in alignment with the rising user's preference. In the subsequent experiments, we generate eleven results by changing the user's preference from 0 to 1 in increments of 0.1 and then calculate the PRS.

6 Experiments

6.1 Experimental Settings

Experimental Tasks

We evaluate the proposed method on four CDR-CA tasks. The first two tasks are diversity increasing tasks: 1) increasing diversity in both appearance and shooting time, and 2) increasing diversity in both appearance and shooting location. The latter two tasks involve mixed objectives: 3) increasing diversity in appearance while decreasing diversity in shooting time, and 4) increasing diversity in appearance while decreasing diversity in shooting location.

Datasets

We use three datasets, i.e., Visual Genome (VG) [Krishna *et al.*, 2017], Incidents 1M (I1M) [Weber *et al.*, 2020; Weber *et al.*, 2022], and PixelProse (PP) [Singla *et al.*, 2024], as they have enough amount of images and have necessary attributes. We employ VG for tasks involving shooting time, I1M for tasks involving shooting location, and PP for tasks involving shooting time or location. We selected images with the necessary attributes from each dataset. Attributes were obtained from EXIF information [1]. The resulting numbers of images are 649, 22,547, and 25,151 for VG, I1M, and PP, respectively. For text queries, we use fifteen object names in VG, the disaster and shooting locations types annotated in I1M, and the 1,000 randomly sampled dense captions in PP. We use 20% of images in each dataset as validation set, with the remaining images constituting the test set.

Evaluation Metrics

- Overall Metric: Our primary metric is the harmonic mean (HM) of the diversity metric and the retrieval accuracy.

- Diversity Metric: As the diversity metric (DM), we report the harmonic mean of the Vendi scores ($VS_{0.1}$) [Friedman and Dieng, 2023; Pasarkar and Dieng, 2024] on each attribute. $VS_{0.1}$ is a metric that evaluates the diversity of each attribute in the top-K result. The higher Vendi score indicates

[1] Please refer to Sec. 3.1 in the supplementary material for the detailed dataset descriptions and preprocessing.

			VG (Time)	I1M (Location)	PP (Time)	PP (Location)
	App.	Att.	HM↑ (MAP↑,DM↑)	HM↑ (MAP↑,DM↑)	HM↑(N@10↑,DM↑)	HM↑ (N@10↑,DM↑)
BLIP-2	-	-	0.8450 (0.8259, 0.8651)	0.7948 (0.7917, 0.7979)	0.7244 (0.6214, 0.8683)	0.7181 (0.6214, 0.8503)
Clustering	✓	-	0.8497 (0.8156, 0.8868)	0.7870 (0.7710, 0.8038)	0.6970 (0.5781, 0.8775)	0.6938 (0.5781, 0.8675)
MMR	✓	-	0.8467 (0.8222, 0.8726)	0.7922 (0.7848, 0.7998)	0.7226 (0.6176, 0.8704)	0.7167 (0.6176, 0.8535)
k-DPP	✓	-	0.8472 (0.8284, 0.8669)	0.7950 (0.7901, 0.8000)	0.7233 (0.6194, 0.8693)	0.7171 (0.6194, 0.8516)
Clustering	-	✓	<u>0.8557</u> (0.8071, 0.9104)	<u>0.8056</u> (0.7837, 0.8288)	0.7226 (0.6086, 0.8891)	0.7140 (0.5967, 0.8888)
MMR	-	✓	0.8475 (0.8282, 0.8676)	0.7968 (0.7879, 0.8058)	0.7207 (0.6108, 0.8787)	0.7123 (0.6025, 0.8710)
k-DPP	-	✓	0.8510 (0.7930, 0.9183)	0.7870 (0.7122, 0.8793)	0.7265 (0.6205, 0.8763)	<u>0.7199</u> (0.6199, 0.8583)
Clustering	✓	✓	0.8422 (0.7984, 0.8910)	0.7981 (0.7799, 0.8171)	0.7204 (0.6060, 0.8878)	0.7096 (0.5916, 0.8864)
MMR	✓	✓	0.8464 (0.8246, 0.8694)	0.7964 (0.7887, 0.8044)	0.7151 (0.5999, 0.8850)	0.7136 (0.6051, 0.8695)
k-DPP	✓	✓	0.8534 (0.7979, 0.9171)	0.7758 (0.7042, 0.8636)	<u>0.7266</u> (0.6208, 0.8759)	0.7197 (0.6200, 0.8575)
MS-DPP	✓	✓	**0.8649** (0.8220, 0.9124)	**0.8081** (0.7618, 0.8604)	**0.7269** (0.6183, 0.8818)	**0.7234** (0.6193, 0.8695)

Table 2: Comparative results with baselines on diversity increasing tasks. The best and second-best are highlighted in bold and underlined, respectively. The first row specifies the dataset name and target attributes. The second and third columns indicate whether appearance (App.) and attribute (Att.) information are used as inputs for each method.

			VG (Time)	I1M (Location)
	App.	Att.	HM↑ (MAP↑,DM↑)	HM↑ (MAP↑,DM↑)
BLIP-2	-	-	0.4200 (0.8259, 0.2816)	0.5654 (0.7917, 0.4398)
Clustering	✓	-	0.4169 (0.6867, 0.2993)	0.5564 (0.7449, 0.4441)
MMR	✓	-	0.4578 (0.8158, 0.3181)	0.5689 (0.8025, 0.4406)
k-DPP	✓	-	0.4198 (0.8284, 0.2811)	0.5637 (0.7901, 0.4382)
Clustering	-	✓	0.5182 (0.6076, 0.4518)	0.5380 (0.7159, 0.4309)
MMR	-	✓	0.5304 (0.7323, 0.4158)	0.6289 (0.7694, 0.5317)
k-DPP	-	✓	0.4039 (0.7335, 0.2787)	0.5515 (0.7397, 0.4397)
Clustering	✓	✓	0.5137 (0.6027, 0.4476)	0.5532 (0.7465, 0.4394)
MMR	✓	✓	<u>0.5311</u> (0.7495, 0.4113)	<u>0.6301</u> (0.7782, 0.5294)
k-DPP	✓	✓	0.4917 (0.6567, 0.3930)	0.5995 (0.6741, 0.5397)
MS-DPP	✓	✓	**0.5746** (0.6616, 0.5079)	**0.6393** (0.7598, 0.5518)

Table 3: Comparative results on the tasks of increasing diversity in appearance while decreasing diversity in an attribute. The best and second-best are highlighted in bold and underlined, respectively. The first row shows the dataset name and the target attribute. The second and third columns indicate whether appearance (app.) and attribute (att.) information are used as inputs for each method.

higher diversity. Given that the Vendi score can range from 0 to K, we normalize the score as $VS_{0.1}/K$ for attributes whose diversity is to be increased and $1 - VS_{0.1}/K$ for attributes whose diversity is to be decreased. We set $K = 20$ across all tasks. A higher value of DM indicates better performance due to the normalization.

- Retrieval Metric: For retrieval accuracy, we use two common metrics: 1) Mean Average Precision (MAP) for VG and I1M, and 2) Normalized Cumulative Semantic Score at 10 (N@10) [Biten *et al.*, 2022] for PP. MAP represents the average value of AP at each query, while N@10 represents the ratio of the retrieved semantic score to the ground truth semantic score. MAP is a suitable metric for VG and I1M due to multiple relevant images per query, whereas N@10 is appropriate for PP, which contains only one relevant image per query. A higher value of each metric indicates better performance.

Implementation Details

We used BLIP-2 [Li *et al.*, 2023], a widely used retrieval model, to calculate the relevance score between the query text and each image. We used the BLIP-2 model finetuned with the MSCOCO dataset [Lin *et al.*, 2014], and employed the ITC mechanism of BLIP-2.

We employed BLIP-2s' image features to calculate an appearance similarity. BLIP-2 outputs 32 features per image. We use the average feature $\mathbf{f}_i$ for each image. We defined the appearance similarity as the inverse of the distance $1/(\|\mathbf{f}_i - \mathbf{f}_j\|_2 + 1)$. For shooting times or locations, similarity was also evaluated using the inverse of the distance between embedded features. Shooting time was embedded as a two dimensional vector $(\cos z, \sin z)$, where $z = \frac{t}{24\times 60}\pi, t = \text{hour} \times 60 + \text{minute}$. Shooting location was embedded as a three dimensional vector $(\cos(lat)\cos(lon), \cos(lat)\sin(lon), \sin(lat))$.

We first select the top-200 images by BLIP-2 and then rerank them using the proposed method to refine diversity.

The hyperparameters of the proposed method were tuned by the grid-search algorithm on the validation set in terms of the HM. The trade-off parameter θ between retrieval accuracy and diversity was selected from 0.75 to 0.95 in increments of 0.05. The attribute weights $\{w_i\}$ were selected from 0.1 to 0.9 in increments of 0.2 and normalized to sum to 1. Whether to use TN or not was determined by the grid search algorithm.

Baselines

We benchmark our method against the original results from BLIP-2 and three single-source diversification techniques: Clustering-based method, MMR, and k-DPP. For the diversity increasing tasks, we employ the standard MMR and k-DPP algorithms with average appearance features or embedded attribute features. For the diversity decreasing tasks, we slightly modify MMR and k-DPP; we multiply the similarity of each attribute by -1. As multi-source baselines, we input concatenated features of appearance and attribute features to the Clustering-based method and MMR. Attribute weights w_i and task indicator s_i are multiplied to feature vectors before the concatenation. As a multi-source DPP baseline, we input the average weighted similarity matrix to k-DPP.

The hyperparameters of the baselines were also tuned using the same strategy as the proposed method. Please refer to Sec. 3.2 in the supplementary material for the other detailed implementation of the baselines.

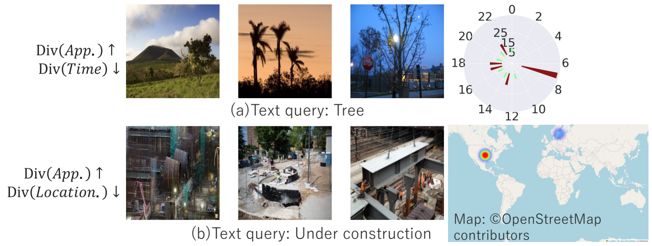

Figure 5: Examples of the results by MS-DPPs. The upper figures show shooting time refinement results, including the top three images and polar histograms of shooting times in the top 20 images. The upward bar of the histogram is at 0:00, and time moves clockwise, rounding the clock in 24 hours. The right figures show location refinement results, including heatmaps of the locations in the top 20 images.

	TVs	M	VG (Time)	I1M (Location)
MS-DPP	-	-	**0.8649** (0.8220, 0.9124)	**0.8081** (0.7618, 0.8604)
	✓	-	<u>0.8557</u> (0.8211, 0.8934)	0.7935 (0.7649, 0.8244)
	✓	✓	0.8425 (0.7893, 0.9034)	<u>0.7938</u> (0.7651, 0.8247)

(a)

	TVs	M	VG (Time)	I1M (Location)
MS-DPP	-	-	**0.5746** (0.6616, 0.5079)	0.6351 (0.7577, 0.5466)
	✓	-	<u>0.5574</u> (0.6687, 0.4778)	<u>0.6373</u> (0.7650, 0.5462)
	✓	✓	0.5558 (0.6678, 0.4759)	**0.6393** (0.7598, 0.5518)

(b)

Table 4: Ablation study on TN for the tasks of (a) increasing diversities of all attributes and (b) increasing diversity in appearance while decreasing diversity in time or location. The second and third columns indicate whether TN is applied for tangent vectors of similarity matrices $\mathbf{S}_i$ and the unified similarity matrix $\mathbf{M}$, respectively.

	VG (Time)		I1M (Location)	
	Inc.	Dic.	Inc.	Dic.
Clustering	1.0876	0.8445	0.7180	0.4514
MMR	-0.3615	0.0873	-0.2449	0.1192
k-DPP	1.3000	-0.1115	0.0171	-0.0561
MS-DPP	-0.2842	0.7341	-0.7272	0.4821
+ TN (TVs)	<u>1.3586</u>	<u>1.3358</u>	<u>1.2849</u>	<u>1.1936</u>
+ TN (TVs + M)	**1.4073**	**1.4196**	**1.2955**	**1.2642**

Table 5: PRS (9) on diversity increase (Inc.) and decreasing (Dec.) tasks.

6.2 Comparative Results

Results for Diversity Increasing Tasks

Table 2 shows the comparative results of MS-DPP and the baselines for the diversity increasing tasks. Our MS-DPP outperforms the baselines in HM for all tasks, demonstrating its capability to enhance the diversity of multiple attributes while preserving retrieval accuracy. This result supports the effectiveness of the MS-DPPs' formulation, which is based on the product of multiple DPP models and the aggregation of multiple attributes in the tangent space of the SPD manifold.

The conventional single-source baselines relying on appearance features alone demonstrate comparatively low performance. This underscores the insufficiency of considering only image appearance for CDR-CA tasks, thus highlighting the necessity of integrating multiple attributes.

Furthermore, our method outperforms multi-source baselines, which show only marginal improvements over their single-source counterparts. This indicates that merely aggregating multiple attributes is inadequate for CDR-CA tasks, whereas our method effectively aggregates multiple attributes for better performance.

Results for Diversity Decreasing Tasks

Table 3 displays the results of our method and baselines for tasks focused on increasing appearance diversity while reducing diversity in a specific attribute. Consistent with previous findings, our method demonstrates superior performance, confirming its efficacy. Notably, there is a significant improvement in diversity scores, supporting the effectiveness of our attribute integration approach in these tasks.

Figure 5 shows the results of two CDR-CA tasks by MS-DPPs. The results demonstrate that our method can effectively refine the diversity of multiple attributes while maintaining highly relevant images [2].

Effect of Tangent Normalization

Tables 4 and 5 show the results of the proposed method with and without TN. Although the overall performance is not always improved, PRS is significantly improved by TN, as shown in Table 5. High PRS is crucial for practical applications, enabling users to adjust each attribute's importance interactively. This result supports the effectiveness of TN in improving the reflection of user preferences, which is the primary motivation for introducing TN.

7 Conclusion

We proposed a novel task, Contextual Diversity Refinement of Composite Attributes (CDR-CA), which is a generalization of the standard result diversification task. CDR-CA targets multiple attributes of each image and has mixed objectives, i.e., increasing diversity in some attributes while decreasing diversity in others. CDR-CA opens up the possibility of addressing various practical applications, unlike the standard result diversification task. To address CDR-CA, we proposed a tailored method, Multi-Source Determinantal Point Processes (MS-DPPs). The basis of MS-DPP is to consider the product of multiple DPP models. We unified the basic formulation into a single DPP model through operations on the SPD manifold. This unification enables to employ the established optimization techniques from the standard DPPs and motivates the introduction of Tangent Normalization (TN), a novel technique that aligns attribute diversities with user-defined weights. Our experiments demonstrated the effectiveness of MS-DPPs in enhancing the diversity of multiple attributes while preserving highly relevant images.

[2]The supplementary material includes additional examples.

References

[Affandi *et al.*, 2014] Raja Hafiz Affandi, Emily Fox, Ryan Aerdams, and Ben Taskar. Learning the Parameters of Determinantal Point Process Kernels. In *ICML*, volume 32, pages 1224–1232. PMLR, 22–24 Jun 2014.

[Benavent *et al.*, 2019] Xaro Benavent, Angel Castellanos, Esther de Ves, Ana García-Serrano, and Juan Cigarrán. FCA-based knowledge representation and local generalized linear models to address relevance and diversity in diverse social images. *Future Generation Computer Systems*, 100:250–265, November 2019.

[Biten *et al.*, 2022] Ali Furkan Biten, Andrés Mafla, Lluís Gómez, and Dimosthenis Karatzas. Is an Image Worth Five Sentences? A New Look Into Semantics for Image-Text Matching. In *WACV*, pages 1391–1400, January 2022.

[Boteanu *et al.*, 2015] Bogdan Boteanu, Ionuţ Mironică, and Bogdan Ionescu. Hierarchical clustering pseudo-relevance feedback for social image search result diversification. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2015.

[Boteanu *et al.*, 2017] Bogdan Boteanu, Ionuţ Mironică, and Bogdan Ionescu. Pseudo-relevance feedback diversification of social image retrieval results. *Multimedia Tools and Applications*, 76(9):11889–11916, May 2017.

[Bouhlel *et al.*, 2017] Noura Bouhlel, Ghada Feki, Anis Ben Ammar, and Chokri Ben Amar. A Hypergraph-Based Reranking Model for Retrieving Diverse Social Images. In *Computer Analysis of Images and Patterns*, pages 279–291. Springer International Publishing, 2017.

[Cao *et al.*, 2022] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-Text Retrieval: A Survey on Recent Research and Development. In *IJCAI*, pages 5410–5417, 2022.

[Carbonell and Goldstein, 1998] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, page 335–336. Association for Computing Machinery, 1998.

[Chen *et al.*, 2018] Laming Chen, Guoxin Zhang, and Hanning Zhou. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *NeurIPS*, pages 5627–5638, December 2018.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024.

[Faghri *et al.*, 2017] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*, page , 2017.

[Friedman and Dieng, 2023] Dan Friedman and Adji Bousso Dieng. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *TMLR*, 2023.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, volume 26, 2013.

[Ionescu *et al.*, 2016] Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. Result Diversification in Social Image Retrieval: A Benchmarking Framework. *Multimedia Tools and Applications*, 75(2):1301–1331, January 2016.

[Ionescu *et al.*, 2021] Bogdan Ionescu, Maia Rohm, Bogdan Boteanu, Alexandru Lucian Gînscă, Mihai Lupu, and Henning Müller. Benchmarking Image Retrieval Diversification Techniques for Social Media. *IEEE TMM*, 23:677–691, 2021.

[Kaur *et al.*, 2021] Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. Comparative Analysis on Cross-Modal Information Retrieval: A Review. *Computer Science Review*, 39:100336, 2021.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017.

[Kulesza and Taskar, 2011] Alex Kulesza and Ben Taskar. K-DPPs: Fixed-size determinantal point processes. In *ICML*, pages 1193–1200, June 2011.

[Kulesza and Taskar, 2012] Alex Kulesza and Ben Taskar. Determinantal Point Processes for Machine Learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, December 2012.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked Cross Attention for Image-Text Matching. In *ECCV*, pages 212–228, 2018.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, pages 9694–9705, 2021.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, pages 12888–12900, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, pages 19730–19742, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Mariet *et al.*, 2019] Zelda Mariet, Mike Gartrell, and Suvrit Sra. Learning determinantal point processes by corrective negative sampling. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 2251–2260. PMLR, 16–18 Apr 2019.

[Odile, 1975] Macchi Odile. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(01):83–122, 03 1975.

[Pasarkar and Dieng, 2024] Amey P Pasarkar and Adji Bousso Dieng. Cousins Of The Vendi Score: A Family Of Similarity-Based Diversity Metrics For Science And Machine Learning. In *AISTATS*, pages 3808–3816. PMLR, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pages 8748–8763, 2021.

[Singla *et al.*, 2024] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From Pixels to Prose: A Large Dataset of Dense Image Captions. In *arXiv:2406.10328*, 2024.

[Sogi *et al.*, 2024a] Naoya Sogi, Takashi Shibata, and Makoto Terao. Object-aware query perturbation for cross-modal image-text retrieval. In *ECCV*, pages 447–464. Springer, 2024.

[Sogi *et al.*, 2024b] Naoya Sogi, Takashi Shibata, Makoto Terao, Kenta Senzaki, Masahiro Tani, and Royston Rodrigues. Disaster damage visualization by vlm-based interactive image retrieval and cross-view image geolocalization. In *IGARSS*, pages 1746–1749. IEEE, 2024.

[Tekli, 2022] Joe Tekli. An overview of cluster-based image search result organization: Background, techniques, and ongoing challenges. *Knowledge and Information Systems*, 64(3):589–642, March 2022.

[van Leuken *et al.*, 2009] Reinier H. van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. Visual diversification of image search results. In *WWW*, pages 341–350, New York, NY, USA, April 2009. Association for Computing Machinery.

[Vieira *et al.*, 2011] Marcos R. Vieira, Humberto L. Razente, Maria C. N. Barioni, Marios Hadjieleftheriou, Divesh Srivastava, Caetano Traina, and Vassilis J. Tsotras. On query result diversification. In *ICDE*, pages 1163–1174, April 2011.

[Wang *et al.*, 2023] Wen Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiangbo Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. *CVPR*, pages 19175–19186, 2023.

[Weber *et al.*, 2020] Ethan Weber, Nuria Marzo, Dim P. Papadopoulos, Aritro Biswas, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. Detecting natural disasters, damage, and incidents in the wild. In *ECCV*, August 2020.

[Weber *et al.*, 2022] Ethan Weber, Dim P Papadopoulos, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. Incidents1M: a large-scale dataset of images with natural disasters, damage, and incidents. *IEEE TPAMI*, 45(4):4768–4781, 2022.

[Wei *et al.*, 2024] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *ECCV*, pages 387–404. Springer, 2024.

[Wu *et al.*, 2024] Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Bowei He, Bhaskar Mitra, and Xue Liu. Result Diversification in Search and Recommendation: A Survey. *IEEE TKDE*, pages 1–20, 2024.

[Yu *et al.*, 2022] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022.

[Zeng *et al.*, 2024] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. $X^2$2-VLM: All-in-One Pre-Trained Model for Vision-Language Tasks. *IEEE TPAMI*, 46(5):3156–3168, May 2024.

[Zhang *et al.*, 2005] Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. Improving web search results using affinity graph. In *SIGIR*, pages 504–511. Association for Computing Machinery, August 2005.

[Zhao *et al.*, 2023] Minyi Zhao, Jinpeng Wang, Dongliang Liao, Yiru Wang, Huanzhong Duan, and Shuigeng Zhou. Keyword-Based Diverse Image Retrieval by Semantics-aware Contrastive Learning and Transformer. In *SIGIR*, pages 1262–1272, New York, NY, USA, July 2023. Association for Computing Machinery.