ModalChorus: Visual Probing and Alignment of Multi-modal Embeddings via Modal Fusion Map

Yilin Ye, Shishi Xiao, Xingchen Zeng, and Wei Zeng

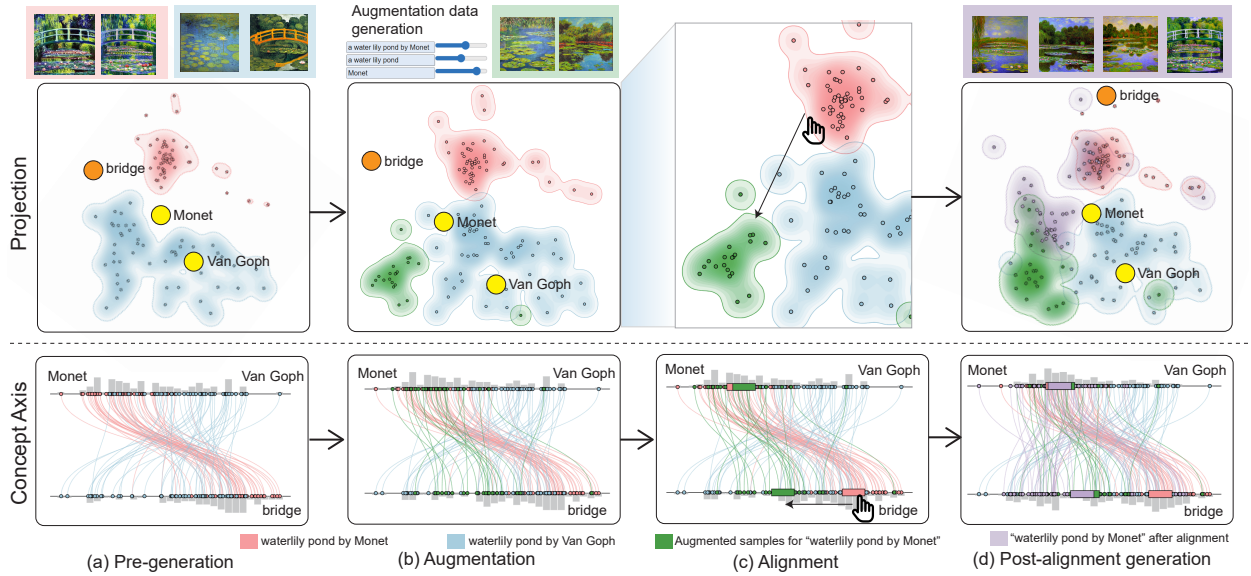


Fig. 1: Visual probing of multi-modal CLIP embeddings for text-to-image generation. (a) For the prompt of "waterlily pond by Monet", users first discover misalignment of pre-trained models in the form of concept entanglement between *"Monet"* and *"bridge"* using our Modal Fusion Map projection and concept axis view. (b) Data augmentation based on weighted embedding generation can be performed to provide extra alignment reference set. (c) Set-set alignment interaction is performed to align the initial generated images of "waterlily pond by Monet" to the augmented images reflecting user intents. (d) Post-alignment model can generate a set of images with more diversity by disentangling *"Monet"* and *"bridge"*.

Abstract— Multi-modal embeddings form the foundation for vision-language models, such as CLIP embeddings, the most widely used text-image embeddings. However, these embeddings are vulnerable to subtle misalignment of cross-modal features, resulting in decreased model performance and diminished generalization. To address this problem, we design *ModalChorus*, an interactive system for visual probing and alignment of multi-modal embeddings. *ModalChorus* primarily offers a two-stage process: 1) embedding probing with Modal Fusion Map (MFM), a novel parametric dimensionality reduction method that integrates both metric and nonmetric objectives to enhance modality fusion; and 2) embedding alignment that allows users to interactively articulate intentions for both point-set and set-set alignments. Quantitative and qualitative comparisons for CLIP embeddings with existing dimensionality reduction (*e.g.*, t-SNE and MDS) and data fusion (*e.g.*, data context map) methods demonstrate the advantages of *MFM* in showcasing cross-modal features over common vision-language datasets. Case studies reveal that *ModalChorus* can facilitate intuitive discovery of misalignment and efficient re-alignment in scenarios ranging from zero-shot classification to cross-modal retrieval and generation.

Index Terms—Multi-modal embeddings, dimensionality reduction, data fusion, interactive alignment

1 Introduction

Neural embeddings are high-dimensional latent representations for knowledge captured from self-supervised pre-training, such as word embeddings and image embeddings. Recently, multi-modal (*e.g.*, text and image) embedding are playing a pivotal role for advancing multi-modal AI models. This type of embeddings learns a joint representation space that encodes different modalities and their relationships, forming the basis for cross-modal tasks such as text-to-image retrieval and generation [3,53,72,74]. The performance of multi-modal embedding models rely heavily on the quality of multi-modal alignment, which seeks to match data with corresponding semantics across different modalities within the embedding space [23, 24, 52]. However, misalignment in multi-modal embeddings is common due to the intricate many-to-many mapping among concepts in different modalities. For instance, text-to-image embeddings can easily encounter misalignment issues of concept entanglement. As illustrated in Figure 1(a), the text prompt of '*waterlily pond by Monet*' becomes entangled with the '*bridge*' concept in the image modality, reducing the diversity of generated images.

Identifying misalignment in multi-modal embeddings is crucial for enhancing model performance. Existing methods for evaluating misalignment often rely on reference-based evaluations (*e.g.*, CIDEr [62] and SPICE [2]) that necessitate extensive human-labeled references, or reference-free metrics (*e.g.*, CLIPScore [32]) derived from pretained multi-modal models. Despite not relying on references, reference-free

- *Y.Ye, S. Xiao, X. Zeng, and W. Zeng are with the Hong Kong University of Science and Technology (Guangzhou) . E-mail: {yyebd@connect., sxiao713 @connect., xzeng159@connect., weizeng@}hkust-gz.edu.cn. Y. Ye and W. Zeng are also with the Hong Kong University of Science and Technology.*
- *Wei Zeng is the corresponding author.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

metrics are reliant on pretrained models, making it challenging for fully automatic methods to detect misalignment in diverse and context-dependent scenarios. For instance, the CLIPScore for the text prompt '*waterlily pond by Monet*' fails to reflect the issue of concept entanglement, as the CLIP model itself is biased towards the '*bridge*' concept in the image modality. Hence, existing fine-tuning techniques to improve alignment often fall short of expectations in numerous scenarios. There is a need for an interactive visualization tool to help users intuitively investigate and address misalignment.

However, the intricate data structures and feature characteristics inherent in multi-modal embeddings pose particular challenges for visual probing and interactive alignment. A key challenge arises from the modality gap, wherein embedding vectors from different modalities are essentially disjointed in the joint embedding space [43]. To achieve cohesive visualization of multi-modal embeddings, it is essential to address the modality gap issue and unify the presentation of different modalities within a single display space. Previous visualizations of neural embeddings have primarily centered on single-modal embeddings, such as word or image embeddings [30, 45, 46]. Notably, these works commonly employ classical dimensionality reduction (DR) methods like t-SNE [61] and MDS [5], which are limited to separately displaying multi-modal embeddings in distinct spaces. Fusion-based DR methods (*e.g.*, [10, 13]) offer a potential solution to jointly project embeddings from different modalities. However, these methods typically treat intra- and inter-modal distances equally, without giving special consideration to cross-modal relations. For example, Data Context Map (DCM) [10] solely relies on metric-based objectives that poorly capture the relative rank order of inter-modal distances. As illustrated in Figure 5(left), DCM projects rather even distribution of image embedding points around the textual concepts, making it harder to observe differences in distribution pattern.

Moreover, enabling interactive alignment for multi-modal embeddings presents another challenge, primarily due to two reasons. First, user-intended alignment strategies encompass diverse operations. For example, in Figure 1, upon identifying the concept entanglement between '*Monet*' in the text modality and '*bridge*' in the image modality, users may prefer to drag the '*bridge*' point far away or relocate the entire set of '*Monet*' images. However, existing studies often focus on point-based operations [17, 68], while others solely support set-based interaction [20]. Secondly, users would utilize interactive alignment to refine the underlying models and ensure that the refined model performs as expected, as illustrated by the disentangled images generated post-alignment, as in Figure 5(d). Existing studies on DR refinement mostly focus on adapting the projection layout [17, 20, 63, 68], whilst overlooking the model refinement.

To fill the gap, we present *ModalChorus*, an interactive system that supports visual probing and alignment of multi-modal embeddings. *ModalChorus* mainly comprises two-stage exploration. First, in the *embedding probing* stage, we propose Modal Fusion Map (MFM), a novel parametric DR method integrating metric and non-metric objectives for enhanced modality fusion. By taking the advantages of metric-based objectives in preserving the intra-modal distances and non-metric-based objectives in capturing inter-modal distance rank order [5, 13], MFM effectively addresses the modality gap challenge induced by multi-modal embeddings. Compared with conventional single-modal and fusion-based DR methods, MFM achieves higher trustworthiness and continuity regarding inter-modal relations (see Table 1), and can better visually reflect the intra- and inter-modality contextual distributions (see Figures 4 & 5). Next, in the *embedding alignment* stage, to accommodate the diverse alignment scenarios, we design an alignment interaction scheme that allows for alignment on multiple levels including point, subset, and set. The interaction scheme is integrated with MFM encompassing point-set and set-set alignment. Besides, a concept axis view is also developed to enable linear visual representation for the probing and alignment of multi-modal embeddings.

In summary, our make the following contributions:

- We propose Modal Fusion Map (MFM), a novel dimensionality reduction method tailored for fusion projection of multi-modal embeddings. The effectiveness of MFM is demonstrated using both quantitative and qualitative evaluations.
- We develop *ModalChorus*, an interactive system that supports visual probing of multi-modal embeddings to discover misalignment, along with an interaction scheme that supports interactive fine-tuning of the underlying multi-modal embedding models.
- We show the effectiveness of our system through case studies on three embedding-based cross-modal tasks, ranging from zero-shot classification to cross-modal retrieval and generation.

2 Related Work

Visualization for Neural Embeddings. Deep learning relies on neural networks that are often pre-trained on large amounts of data. Neural embeddings are the foundational high-dimensional feature representation of raw data encoded by neural networks, such as text embeddings like word2vec [49] and BERT [16] and image embeddings like SimCLR [9]. Visualization researchers have dedicated significant efforts to enhance the comprehension of neural embeddings. Previous studies primarily focus on unimodal embeddings, encompassing word embeddings [30, 31, 45] and image embeddings [46]. Many of these studies integrate projection methods with axis-based [30, 45, 46] or set-based [31] exploration techniques. For example, Liu *et al.* [45] identified analogy axis between multiple pairs of words with the same semantic transition in word embeddings projected by t-SNE. Latent Space Cartography [46] extends the concept of semantic axis to customized axis defined by users, which can be applied to exploration of both unimodal word embeddings and image embeddings. EmbComp [31] combines t-SNE projection with visualization of neighborhood set overlap to compare different word embedding models.

Recently, multi-modal embeddings such as CLIP [52] and ALIGN [37] have fueled the advances in multi-modal AI such as text-to-image generation. These embeddings can encode data from different modalities in a joint space, contributing to various applications such as cross-modal retrieval [4] and generation [53, 77]. However, this integration also introduces modality gap [43] that signifies discrepancies between different modal embeddings, complicating the comprehension of multi-modal embeddings. There is a lack of visualization tool tailored to the task. Specifically, the task demands an effective visualization method for probing multi-modal embeddings and an interactive scheme for improving alignment of multi-modal embeddings. To meet the goal, we propose the Modal Fusion Map that can better preseve the contextual information of multi-modal embeddings, and an interactive alignment scheme that offers visual steering for modal alignment.

Contextual Dimensionality Reduction. Dimensionality reduction for multi-modal data has been a challenging problem as traditional DR methods like t-SNE [59, 61], PCA [64], and MDS [5] cannot account for cross-modal relations due to the modality gap [43, 50, 80]. Contextual visualization is a type of DR method designed to project data points in relation to attribute points [10, 48, 78], which can be applied to multi-modal data projection, yielding more integrated visualization than dual analysis [15]. Existing contextual visualizations can be categorized into two types: anchor-based projection and fusion-based methods. Anchor-based methods employ a two-stage approach, initially determining the layout of points in one modality before calculating the position of points in the other modality. For example, the RadViz method [11, 33, 73] first lays out the attribute points on a circle and then projects the data points based on their multi-dimensional attribute values. However, the structure of the embedding space can be significantly distorted due to the challenge of optimally laying out the anchor points.

One type of fusion methods, known as co-embedding methods [12, 70], introduces their own high-dimensional representations of multi-modal data or modifies the embeddings of certain data points to achieve a desired visual layout. However, these methods diverge from our goal as they alter the original embeddings with custom models, which cannot help users understand commonly used multi-modal embeddings in AI tasks. Other fusion methods are limited to more specific conditions [25, 26, 67], such as COPE [25] which requires co-occurrence statistics. Visualization researchers have developed more general fusion methods [10, 78]. Particularly, Data Context Map (DCM) [10]

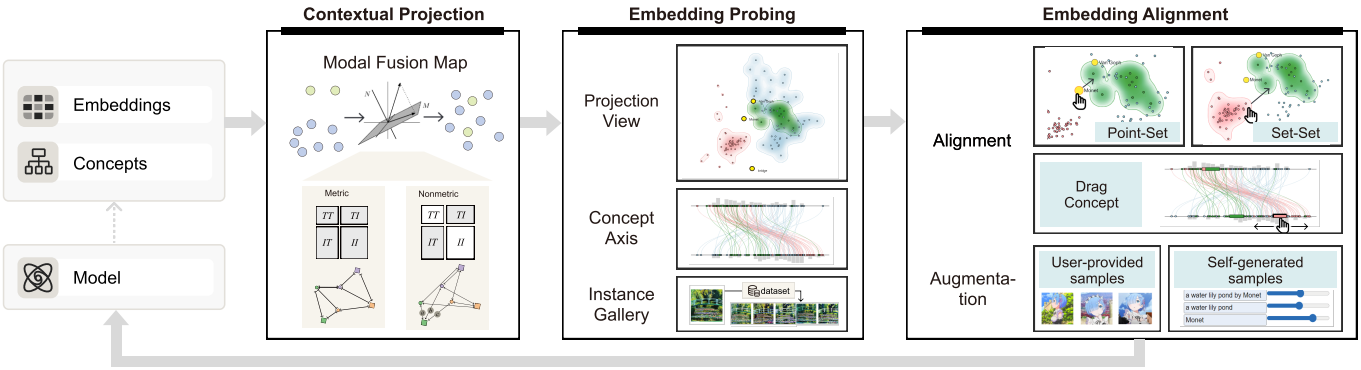


Fig. 2: Overview of our framework. Multi-modal embeddings and concepts extracted from text and images are first projected with Modal Fusion Map, a novel modality-fusing DR method we propose. In the visual exploration stage, visual probing of the embeddings is enabled in the projection view and the concept axis view, allowing users to explore embedding sets and individual instance point. Finally, in the embedding alignment stage, interactive alignment with point-set and set-set alignment schemes is provided, along with optional augmentation with few-shot samples.

defines the distance matrix for the attributes and merges it with the data distance matrix before using MDS to jointly project the attribute points and data points. However, in DCM, intra-modal and inter-modal distances are equally treated in metric-based optimization, which limits its ability to adequately capture the cross-modal non-metric ordinal structure of multi-modal embeddings. In our study, we introduce the Modal Fusion Map, which integrates both metric and nonmetric objectives into modality fusion using a novel parametric DR method, to effectively preserve relationships for intra- and inter-modal distances of multi-modal embeddings.

Visual Steering for Modal Alignment. Pre-training of multi-modal foundation embedding models relies on alignment through methods like contrastive learning, such as ViLBERT [47] and CLIP [52]. For example, the CLIP embeddings [52] is pre-trained on large-scale image-text pair corpus by matching image and text caption in a joint representation space with contrastive multi-class N-pair loss. The pre-training methods typically aim to establish a foundational model, yet the varying quality of pre-training data often leaves some misalignment in specific cases, which requires adaptation such as few-shot fine-tuning [21,35,50] to refine the alignment. Misalignment cases may require human knowledge to be discovered, and the fine-tuning process also typically involves users' choice of alignment data and direction, which necessitates an interactive system to support human-in-the-loop workflow. This scenario differs from interactive prompt engineering of pre-trained models [19,28,60,66], as prompt engineering only seeks to alter the input without refining the model, which is not enough for steering complex multi-modal models with misalignment.

Some previous visual analytics systems support interactive improvement of AI models through label correction or data augmentation [7,27,29]. For example, VATLD [27] leverages disentangled representation learning for semantic exploration of traffic light detection results in relation to explainable data dimensions. However, these studies only focus on task-specific models without paying attention to foundational embeddings [71]. Many studies also rely on the ground-truth labels for insight discovery, which may not be available in real-time probing of pre-trained models. In addition, these studies lack support for visual steering interaction directly in the visualization space, which is more intuitive for the alignment operation our study aims at.

Some visualization researchers have studied interactive visual steering of dimensionality reduction results [17,20,63,68]. For example, Xia *et al.* [68] proposed a contrastive learning-powered parametric dimension reduction method to support point-level interaction to enhance the visual clustering effect. ULCA [20] supports set-level visual steering interaction for comparative analysis. DRAVA [63] introduces an interaction method to adjust the positions of small multiples in axis-based visualization based on βVAE. However, these interactions only focus on refining the projection layout for visual exploration purposes, lacking the ability to align the underlying models or high-dimensional representations. In addition, the interaction schemes of most previous studies are limited to a single type of interaction, such as point-based or set-based interaction in a single view, which cannot cover the diverse alignment scenarios of multi-modal embeddings. In our study, we develop an interaction scheme supporting point-set and set-set alignments, enabling flexible alignment of underlying embedding-based models.

3 Overview

3.1 Background and Domain Problem

Multi-modal embedding. Multi-modal embedding models are pre-trained encoder models for the representation of multi-modal data. For example, the CLIP model is pre-trained on a large corpus of image-text pairs, using transformers and vision transformers to first separately encode text and image into high-dimensional vectors. Then, through a linear transformation, the text embedding and image embedding vectors are aligned in a shared embedding space with contrastive loss. Multi-modal embeddings are the foundational encoder for many AI tasks that involve multi-modal data in its input and/or output. Common tasks include semantics-based image classification [52], cross-modal retrieval [24], and text-to-image generation [8,54].

Alignment. In multi-modal models, alignment means the matching of data representations with corresponding semantics from different modalities. In the pre-training stage of CLIP, for example, the alignment is achieved by updating the embeddings of an image and its corresponding text caption so that they are closer than incorrect pairs in the high-dimensional representation space. However, due to the varying quality and large quantity of data in pre-training and imperfection in training algorithms, there may be misalignment in the pretrained model, which requires further adaptation for enhancing alignment [21,35,50].

Multi-level Alignment. There is mainly alignment on three levels: point, subset, and set, requiring two types of alignment: point-set alignment and set-set alignment. Specifically, users may discover misalignment of an individual point (*e.g.*, misclassified image point or misunderstood text point), a subset (*e.g.*, a subset of incorrect samples in the whole set of text-to-image retrieval results), and a set (*e.g.*, biased or entangled generation results of text-to-image models), requiring different alignment operations. To clarify, we refer to keywords extracted by our system or entered by users as concepts, which are the main text embeddings we focus on in this study for contextual exploration of embeddings, while particular image point is referred to as instances.

Problem: visualization for embedding. Many visualization studies treat embedding methods as a tool for processing data, with the aim of optimizing the visual display of raw input data. That is, these visualizations regard embedding as a projection method. Instead, in the field of AI, representation learning of single and multi-modal embeddings has been playing a pivotal role for various downstream tasks [41,55]. The high-dimensional embeddings themselves are the key intermediate representations of data extracted from raw text or pixels, not just the representation for visual display only. To gain insight into large AI models, particularly for the alignment problem, high-dimensional

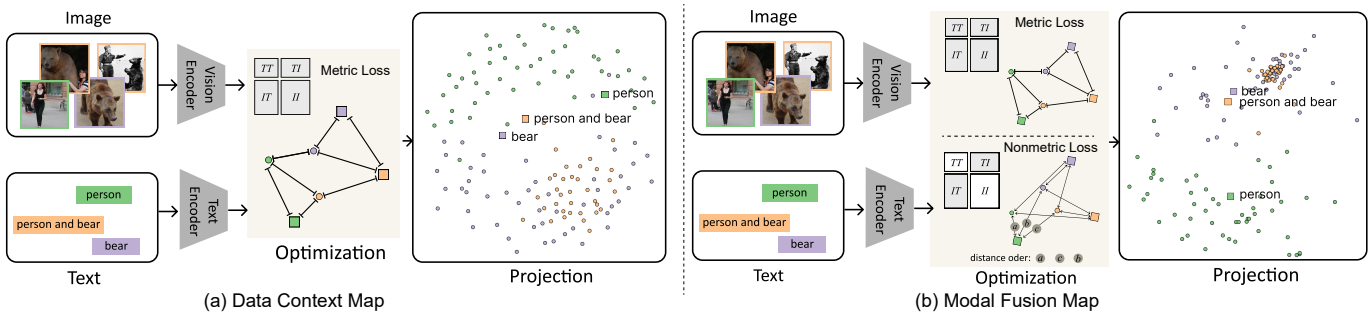


Fig. 3: (a) Data Context Map (DCM) only considers metric-based optimization that indiscriminately seeks to preserve the absolute distance for intra-modal and inter-modal pairs of data points. (b) Inspired by the observation that the nonmetric rank order of cross-modal distances is important for multi-modal embedding-based tasks, our Modal Fusion Map (MFM) combines metric and nonmetric objectives for fusion.

visualization methods should *prioritize capturing the features of the foundational embeddings itself (**G1**)*. For example, suppose two classes of images are indeed close in the embedding space, which signifies risks of misalignment in the embedding. In that case, we do not wish to maximize the class separation in projection space just for visual display since it can mislead users. For another example, if the prompt's text embedding in the generation model is not close to the generated images, we should be cautious about directly putting the text at the centroid of the generated image set, which may lead users to believe that the generation is fully aligned.

To summarize, the former studies focus on embedding for visualization, while our work aims at visualization for embedding. In addition, the aim for interaction in this scenario is to *improve the foundational high-dimensional embeddings (**G2**)* instead of improving the visual display of data like previous studies did [20, 68].

3.2 Challenges and Design Requirement

Accomplishing these two goals is challenging for existing visualization methods, particularly due to:

C1 Modality Gap. The heterogeneous distributions of different modalities in the joint embedding space result in the modality gap [43,80], making it difficult for existing DR methods to simultaneously capture intra-modal and cross-modal features.

C2 Diverse alignment intentions. The diverse alignment scenarios in different cross-modal tasks post challenges to designing a comprehensive interaction scheme integrated into the visualization.

To tackle the challenges, we summarize the design requirements of *ModalChorus*, which should support flexible and effective *R1) visual probing* of multi-modal embeddings to meet ***G1***.

R1.1 Accurately preserving inter- and intra-modal distances. An effective fusion-based DR method is needed to bridge the modality gap while maximally preserving inter- and intra-modal relations.

R1.2 Effective visual presentation to help identify misalignment. Apart from the projection, effective graphical enhancement is needed to assist discovery of misalignment issues such as misclassification or entanglement.

Second, *ModalChorus* shall facilitate *R2) interactive alignment* of multi-modal embeddings to support ***G2***:

R2.1 Supporting alignment on point and set levels. Users may discover embedding misalignment on an individual data point or a whole set of points, demanding different types of alignment interaction, including point-set and set-set alignment.

R2.2 Supporting axis-based alignment. Previous embedding visualization studies have identified the semantic axis as an effective complement of the overall projection for more focused concept-related exploration [30,45,46]. Besides directly manipulating the projection of embeddings, users also need to perform axis-based alignment as the axis can more clearly show the direction of alignment with respect to a specific semantic concept.

R2.3 Supporting data augmentation. When users discover misalignment but cannot find correct reference data, they would like to provide extra data and process it to help the alignment.

3.3 ModalChorus Overview

An overview of our system is shown in Fig. 2, which mainly consists of two stages: 1) embedding probing and 2) embedding alignment. In the first stage, starting from a particular dataset and task, along with user-provided input or automatically extracted concept, we support visual probing of the embeddings with sampled data for interpretation of embeddings and discovery of misalignment. Particularly, we develop Modal Fusion Map, a novel parametric fusion method that integrates metric and nonmetric objectives for multi-modal embedding projection. We also incorporate a concept axis view that allows users to explore the correlation of image embeddings in relation to concept text embeddings. An additional instance gallery displays similar images to the selected image point in the embedding space for neighborhood exploration.

In the second stage, upon discovering misalignment, we enable users to select a particular point, subset, or set and perform point-set alignment or set-set alignment in either the projection view or the concept axis view. In some cases, when new data is needed to enhance the alignment, we allow users to upload their collected data for few-shot alignment or use our system's weighted embedding generation function to generate candidate augmentation data. Finally, the visual alignment operations are mapped to the backend fine-tuning.

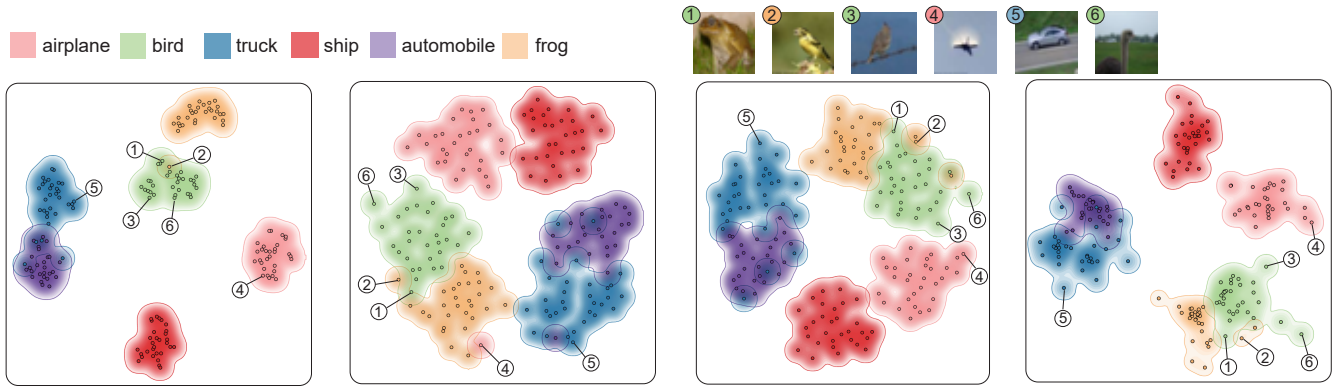
4 Multi-modal Contextual Visualization

In this section, we describe Modal Fusion Map, a novel DR method we propose to address *R1 visual probing* of multi-modal embedding.

4.1 Problem Identification

To address the modality gap problem in multi-modal embedding visualization, data matrix fusion methods [10, 13] are a promising solution. Matrix fusion methods such as Data Context Map (DCM) [10] are derived from the MDS method for distance-based fusion. The original Data Context Map is designed for the attribute and data spaces of multi-dimensional data. Specifically, to align data points from different modalities, it constructs a large distance matrix containing the pairwise distances between all the data points and attribute points, where the intra-modal distance is the original high-dimensional distance such as Euclidean or Cosine distance while the cross-modal distance needs to be defined according to data properties. For example, DCM defines the distance between attribute point and data point as $1 - v$, where v is the data point's value in this attribute dimension.

First, to account for high dimensional latent space, we can naturally change the attribute-data distance in DCM to the Cosine distance between text embedding and image embedding. However, this modification may not suffice for the complexity of multi-modal embedding. Specifically, to enhance the modality merging effect, it is important to flexibly adjust the weights of intra-modality and inter-modality distance. Directly scaling the submatrix as mentioned in [13] may have the risk of significantly distorting the embedding space or exacerbating the modality gap. More importantly, when multi-modal embeddings



(a) t-SNE (b) MDS (c) DCM (d) MFM

Fig. 4: For the zero-shot classification task on CIFAR-10 which relies on cross-modal similarity (color of points represent the predicted class), MFM can better reflect set relations and outliers for visual probing of misalignment.

like CLIP are used for cross-modal tasks such as text-to-image retrieval, the absolute distance between the text and image embeddings is less important than the relative order of the distance values. In visualization of multi-modal embeddings, such characteristics should also be considered. This means that we should develop a better fusion method that considers both metric and non-metric objectives [18, 34, 51].

4.2 Modal Fusion Map

Dimensionality Reduction. Inspired by recent work on parametric dimensionality reduction [68, 69, 75], to satisfy the projection requirements presented above, we propose Modal Fusion Map (MFM) which can flexibly combine different objectives for joint multi-modal embedding projection. The hypothesis is that in the high dimensional embedding space, there exists a subspace or manifold surface S between the text embeddings set T and image embeddings set I, such that the projection of embeddings from both modalities on this surface $(P(T), P(I) \in S)$ can result in an optimized 2D parametric representation $S(x,y)$. Specifically, like other matrix-based methods, we first compute the merged distance matrix:

$$M = \begin{pmatrix} II, & IT \\ TI, & TT \end{pmatrix}, \tag{1}$$

where II is image distance submatrix, TT is text distance submatrix, $IT = TI^T$ is the cross-modal distance submatrix, all using cosine distance or the equivalent Euclidean distance between normalized vectors. Each submatrix is normalized by their mean value.

Next, instead of directly applying the traditional MDS method to the merged matrix as in DCM, we parametrize the projection with a three-layer feed-forward neural network mapping 512 or 1024-dimensional CLIP embedding to the 2-dimensional projection space. See supplementary material for more detail. Then, to implement the MDS objective, we construct a loss function using the Pearson correlation between the high dimensional merged distance matrix and the projected distance matrix for scale-free optimization.

$$L_M = -\frac{\sum(M_{i,j} - \overline{M})(P_{i,j} - \overline{P})}{\sqrt{\sum(M_{i,j} - \overline{M})^2}\sqrt{\sum(P_{i,j} - \overline{P})^2}}, \tag{2}$$

where P stands for the distance matrix of the projected points.

In this way, we can easily define loss terms for the intra-modality and inter-modality submatrices, denoted as L_{TT}, L_{II}, and L_{IT}, respectively. Accordingly, the loss function for metric MDS is the weighted sum. In our case, we only consider the overall term and the cross-modal term: $L_1 = w_1 L_M + w_2 L_{IT}$, where we set $w_1 = 10, w_2 = 2$.

In addition, for the nonmetric loss to preserve cross-modality distance order, we further introduce another loss term:

$$L_2 = \frac{-\sum_{j<k} f((TI_{i,j} - TI_{i,k}) * (P(TI)_{i,j} - P(TI)_{i,k}))}{\|P(TI)\|_2}, \tag{3}$$

Table 1: Evaluation of projection methods with inter-modal and intra-modal trustworthiness (T) and continuity (C) metrics.

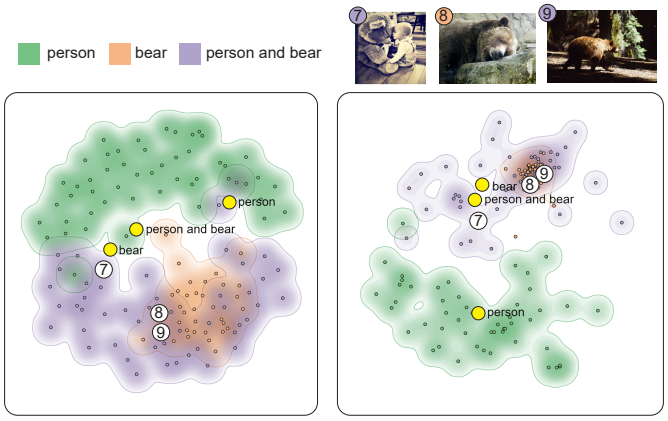
	Inter-modal		Intra-modal	
	T(30)	C(30)	T(30)	C(30)
PCA	0.9177	0.9301	0.7297	0.8183
MDS	0.9274	0.9336	0.8039	0.8537
Isomap	0.9307	0.9281	0.7706	0.8637
t-SNE	0.9290	0.9296	**0.9098**	0.9010
NDCM	0.9223	0.9225	0.5304	0.5309
DCM	0.9385	0.9434	0.8481	0.8941
MFM	**0.9589**	**0.9645**	0.8764	**0.9117**

where $f(x) = \begin{matrix} 0, x \geq 0 \\ -x, x < 0 \end{matrix}$. This loss term will be zero when all the cross-modal distance order is preserved in the projection. The final loss $L = L_1 + \alpha L_2, \alpha = 0.05$. w_1, w_2, α are selected empirically. Code is available at: https://github.com/yilinye/Modal-Fusion-Map.

Contour-based graphical enhancement. We provide graphical enhancements in the form of density contour as inspired by recent work [78]. As shown in Fig. 5, the density plot can show the default KDE density estimation of data point distribution. The KDE contour can serve as a graphical representation of sets in the projection view, which can facilitate subsequent alignment interaction as we describe below. Alternatively, when users provide customized metrics defined for the data points, such as CLIP-Score for generated samples, the density plot can show the kernel estimation of the metric value distribution.

4.3 Evaluation

Qualitative Comparison. As shown in Fig. 4 and Fig. 5, MFM has many advantages for displaying both intra-modality and inter-modality features compared to the DCM method and traditional projection methods like MDS and t-SNE. Specifically, Fig. 4 displays an intra-modal case with the projection of CLIP image embeddings for samples of 6 classes in CIFAR-10 dataset. The colors represent the zero-shot classification results based on CLIP. Among the results, we can see that t-SNE achieves the best separation effect. However, t-SNE also has significant drawbacks in understanding the embeddings and identifying misalignment because it does not consider cross-modal features. First, t-SNE is weaker at showing contextual information, such as the relation between different sets. For example, we can find in Fig. 4, the frog set (green point 1) can be confused with the bird set (yellow point 2) because of similar color or background, yet the t-SNE projection does not clearly show the relation compared to MFM. In addition, our joint projection also shows better within-set distribution than t-SNE. For example, with MFM, we can clearly see outliers or border points within sets (*e.g.*blue point 5 and green point 6). Point 5 corresponds to an image of a car driving on a highway, while most other car images are static scenes of parked cars. Point 6 is a long-necked ostrich that is quite different in appearance from other birds. However, these points are hard to identify



(a) DCM (b) MFM

Fig. 5: DCM insufficiently captures the rank order of cross-modal distances between text and image embeddings, resulting in rather even distribution of image embedding points around the concept text embeddings, making it harder to observe differences in distribution pattern.

in the t-SNE projection. The MDS result in Fig. 4 (b) is more effective than t-SNE for showing the pointwise relationship, but the clustering effect is apparently weaker than t-SNE and MFM. In addition, MDS tends to distribute the points quite evenly in the projected space, which compromises the display of in-set distribution and outliers. The DCM method shows the contextual set relationship better than t-SNE and displays set outliers only slightly better than MDS, since for the image modality, both DCM and MDS use metric loss, but MFM achieves better effect in both aspects. Additionally, we compute Z-Score for a data point to help verify whether a visual outlier or border point is indeed so in the original high-dimensional space (see supplementary for detail). We find that point 6, the most obvious outlier in Fig. 4 (d) indeed has the highest Z-score of 1.2379.

Fig. 5 shows the inter-modal case with the projection of CLIP text embeddings of three text queries together with the image embeddings of the query results. Regarding the DCM results (Fig. 5 (a)), it does successfully merge the modality. However, as shown in both cases, DCM has a significant weakness in that it scatters out the image embedding points quite evenly across the space, making it difficult for users to find distributional differences between different regions. In comparison, for example, in Fig. 5 (b), we can find an obvious dense cluster in MFM results which contain many similar images of bear in the wild, but this pattern is less evident in DCM results. In addition, MFM also more clearly shows the relation between concepts than DCM, as we can find that the query results of *bear* and *person and bear* have obvious overlap in both DCM and MFM results, signifying the closer relationship between these two text concepts, but the relative position of text embeddings in MFM is more coherent to this relation.

Quantitative Evaluation. To show some quantitative evidence of the advantage of MFM, we evaluate the method on COCO dataset using the trustworthiness metric and continuity metric [38, 68], where the former calculates how faithfully the projected kNN reflects the true kNN's in the embedding space and the latter calculates how well the original high-dimensional kNN is preserved in the projection. Particularly, we calculate both inter-modal and intra-modal kNN for k=30. However, in our scenario the inter-modal metric is more important as it measures the methods' ability to preserve multi-modal embedding structure. For the evaluation process, we perform multiple rounds ($r = 500$) of evaluation where in each round we randomly sample 500 images from COCO and project them together with the 80 category text embeddings in COCO object labels. The final metric is the average of the results in all rounds. As shown in Table 1, the experimental results indicate that MFM method performs consistently better than all the other methods in inter-modal truthworthiness and continuity, with higher than 2% margin over the strongest baseline DCM. In addition, MFM also achieves good performance in intra-modal metrics, only second to t-SNE. NDCM [13] is another fusion method using fully nonmetric objective. We can see that among the three fusion methods (MFM, DCM and NDCM), our MFM is consistently better across inter-modal and intra-modal metrics, while fully nonmetric fusion method has significant disadvantage in keeping the intra-modal features. We also need to note that the inter-modal metrics for non-fusion traditional methods like t-SNE and MDS cannot fully reflect their weakness in inter-modal scenarios because the modality gap will cause large distances between image embeddings and text embeddings in the projection space, making it difficult to perceive the differences between the inter-modal distances [43, 50, 80].

5 *ModalChorus* System

5.1 Visualization Interface

Settings Panel. The settings panel (Fig. 6 (a)) allows users to specify some basic settings for their exploration, including tasks and inputs. Users can also select specific concepts in their input to produce contextual visualization in the projection view. Instead of relying solely on textual concepts explicitly extracted from existing text labels or prompts, ModalChorus extracts implicit concepts from images to provide a comprehensive display of concepts. To achieve this, we first leverage BLIP-2 [42], a multi-modal language model capable of receiving images as input and generating textual descriptions of those images. We then employ the TopicRank [6] algorithms to extract candidate visual concepts based on the text generated by BLIP-2.

Projection View. The projection view (Fig. 6 (b)) is the main view of the system leveraging our proposed Modal Fusion Map to help users probe the embedding with different tasks and data. Users can choose to turn on or turn off the contour to emphasize set relation or facilitate instance exploration respectively. The projection view also includes an instance retrieval subview below (Fig. 6 (c)). Users can mouse over the embedding point to see the corresponding image in the gallery. They can also click the point to retrieve similar images to the selected one. In addition, users can select a subset of points by lasso or ctrl-click, as shown in Case 2 and Fig. 10 in Sect. 6.

Concept Axis View. As shown in Fig. 6 (d) and Fig. 7 (a), the concept axis view supports axis-based exploration of image embeddings in relation to text embeddings for user-selected concepts from the settings panel. Users can define one-end axis with a single concept (*e.g.*, bridge) or two-end axis with opposing concepts they want to contrast (*e.g.*, Monet and Van Goph). For one-end axis, the position of an image embedding point x is:

$$\mu_A(x) = l \cdot \frac{sim(x,A) - min(sim(\hat{x},A))}{max(sim(\hat{x},A)) - min(sim(\hat{x},A))}, \tag{4}$$

where $sim(x,A)$ denotes the cosine similarity between x and text embedding of concept A in embedding space, l is the length of the axis. For two-end axis, the position of x is calculated as $l \cdot (0.5 + \frac{\mu_A(x) - \mu_B(x)}{\mu_A(x) + \mu_B(x)})$. When users define more than one axis, we use curves connecting the same instance on two axes to show the correlation. Histogram is also used to help users see the overall distribution. Apart from displaying instances of image embeddings, the concept axis can also represent the whole set or subset as small box at the average position of all the in-set points, showing users the mean value of the set and supporting further set-based alignment interaction as described in Fig. 8 and Sect. 5.2. We also allow users to switch to a scatterplot visualization (Fig. 7 (b)).

Augmentation Panel. The data augmentation panel (Fig. 6 (e)) supports interactive augmentation of alignment data. In some alignment scenarios, users cannot find proper alignment data from the original dataset (for example, users may not find any satisfactory results generated by a pre-trained generative model). For such a problem, the augmentation panel first allows users to upload a subset of samples to supplement the alignment data. For unlabeled raw image data uploaded, this panel also integrates an auto-tagging function based on CLIP-interrogator [1], which can generate tags associated with the image to enhance the alignment performance. Second, in cases where users even find it difficult to collect their own data, the augmentation panel also incorporates a generation function that enables users to leverage the weighted sum of existing text embeddings [14, 65]

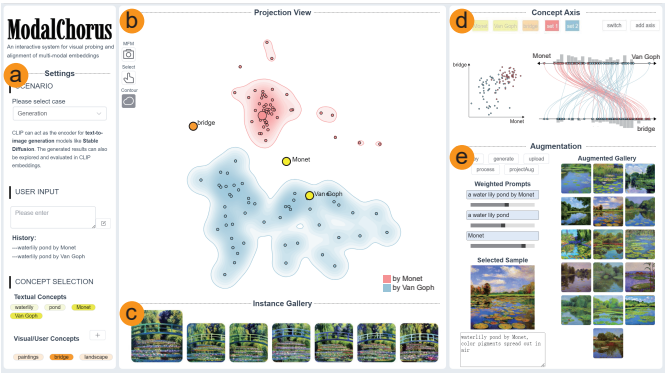


Fig. 6: ModalChorus system. (a) Settings panel on the left allow users' choice of task and dataset. The main projection view (b) displays the MFM dimension reduction result of embeddings. The concept axis view (d) supports axis-based exploration, while the augmentation panel (e) facilitates uploading, generating, and tagging additional data for alignment.

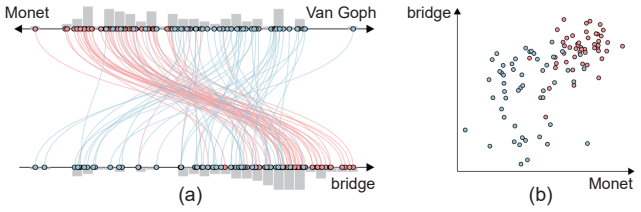


Fig. 7: The concept axis view allows the definition of unidirectional and bidirectional concept axes, showing the distribution of image embedding points in relation to concepts by similarity to concept text embedding. Curves linking data points on different axes display the correlation between the distribution and reveal patterns such as entanglement.

in a pre-trained generation model to synthesize more candidates of intention-aligned samples.

5.2 Interactive Alignment

Alignment Interaction Design. We design a series of visual alignment interactions, which allow users to intuitively express diverse alignment intentions through visual metaphor and trigger backend fine-tuning without writing complex training code. As shown in Fig. 8, our interaction scheme supports set and point level intentions for alignment. First, the common types of alignment mainly concern data points or subsets of points, which we categorize into two types: point-set alignment and set-set alignment, as shown in Fig. 8. First, *point-set* alignment encompasses various scenarios of aligning a set with a data point and vice versa. For example, when users want to align a subset of retrieval results with a query text embedding or when users want to align a prompt embedding to fine-tune samples provided by themselves. As shown in Fig. 8 (a), point-set alignment can be performed on either the projection view or the concept axis. Formally, the high-level idea of point-set alignment can be summarized as follows: Suppose we have a CLIP-based model $F(\cdot)$ which can map input text or image to different sets $C_1, C_2, ..., C_N$ in the sampled data. For example, in classification, C_i corresponds to the set of embeddings for a predicted class; in retrieval and generation, C_i corresponds to the set of embeddings for the results of a single query or prompt. Given a user-selected misaligned image or text point p, the target of point-set alignment is to tune the weights of $F(\cdot)$ such that $\hat{F}(p)$ is closer to the correct set C_i in the embedding space. Although the concrete implementation may vary for different tasks, in terms of the merged distance matrix, the effect is equivalent to achieving the following contrastive objective:

$$\frac{1}{|C_i|}\sum_{v\in C_i} M_{\hat{F}(p),v} < \frac{1}{|C_j|}\sum_{u\in C_j} M_{\hat{F}(p),u} \quad \forall j \neq i, \tag{5}$$

where the estimated distance between $F(p)$ and C_i should be smaller than any other set C_j. Second, *set-set* alignment involves moving two

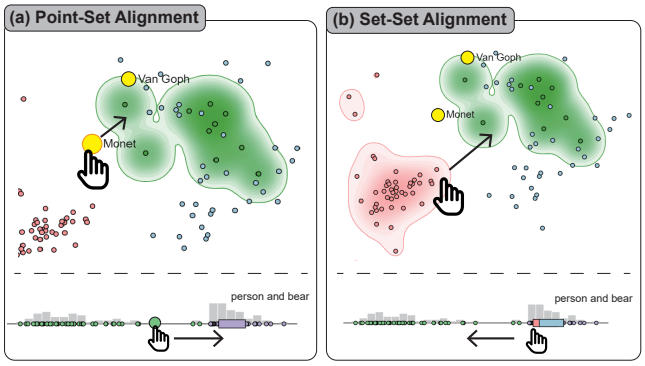


Fig. 8: We design interactions that allow users to visually express their alignment intentions, including point-set alignment and set-set alignment performed in both projection view and concept axis view.

subsets of points closer or further in the concept axis or projection view. Such alignment is intended to close the gap between two sets or distributions in the embedding space, or contrast two sets for distinguishing them better, which can be useful for cases like merging or disentangling concepts in retrieval or generation. As shown in Fig. 8 (b), in the projection view, they can drag a set contour towards another, while in the axis view, they can drag one set box closer to or away from another. Formally, the high-level idea of set-set alignment is: Suppose users identify a misaligned set or subset of embeddings C_e, where C_e is not align with the input p. Next, users find another correct set C_i either by visual exploration of other projected data points or by data augmentation. The goal of set-set alignment can then be formulated as:

$$\frac{1}{|\hat{F}(p)|\cdot|C_i|}\sum_{v\in\hat{F}(p),u\in C_i} M_{u,v} < \frac{1}{|\hat{F}(p)|\cdot|C_e|}\sum_{v\in\hat{F}(p),u\in C_e} M_{u,v}. \tag{6}$$

Alignment Fine-tuning Implementation. Our system provides a general framework to map users' visual interactions shown in Fig. 8 to backend fine-tuning operations that align the model's output in the embedding space. As the visual representations are decoupled from actual backend implementation, our framework can incorporate any kind of specific fine-tuning methods. For demonstration purposes, our study implements two methods. First, for the classification and retrieval cases, we implement triplet loss [58] based alignment. Second, for the generation cases, we implement the low-rank adaptation method [35]. More detail is provided in the supplementary material.

6 Case Studies

In this section, we perform three case studies to demonstrate the usefulness of the Modal Fusion Map and ModalChorus system, which cover three different tasks based on multi-modal embeddings, including zero-shot classification, text-to-image retrieval, and generation. Particularly, we demonstrate how our visual probing integrates with and enhances interactive few-shot alignment [22, 36].

6.1 Case 1: Zero-shot classification

In this case, we demonstrate how our system can be used to visualize the zero-shot classification [52] based on multi-modal embedding clustering and help refine the embedding interactively with one-shot point-set alignment. Specifically, we use the CIFAR-10 image classification dataset [40] to show an example. Here we suppose no ground-truth labels are available. This is to simulate real-time analysis of zero-shot embedding-based classification in the wild for unknown data, where interactive visual analysis with human intervention is most helpful.

Users first select the classification task and the dataset. Then, users subjectively select some classes that they suspect may be confusing for CLIP, including classes of small wild animals and classes of vehicles. Specifically, they select 6 class concepts they want to explore, including *airplane*, *automobile*, *truck*, *ship*, *bird* and *frog*. Then, the system

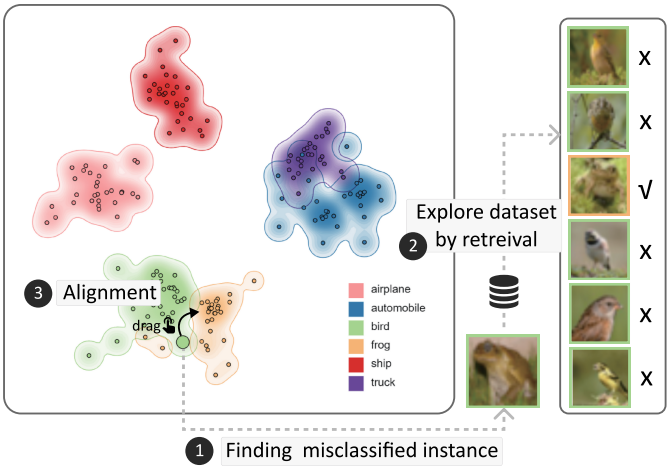


Fig. 9: In zero-shot classification with CLIP, users can leverage *ModalChorus* to identify potential examples of misalignment (in this case, classification mistake) and perform point-set alignment to refine CLIP.

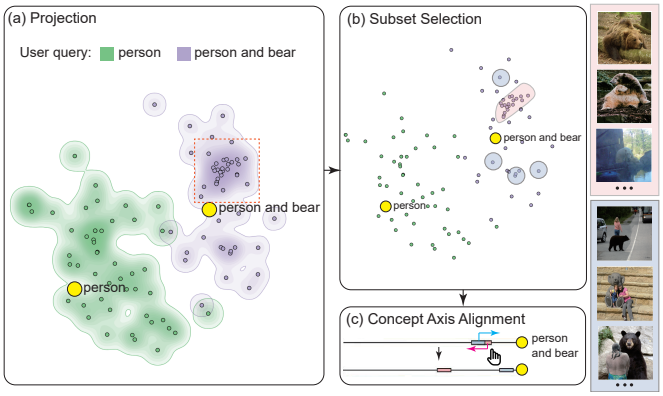


Fig. 10: In cross-modal retrieval with CLIP, users can leverage *ModalChorus* to visualize both the query text and image results embeddings. Users can further select subsets of data and perform axis-based alignment.

predicts the class of each image according to the cross-modal proximity between the image embedding and class text embedding. The system samples the data for visual probing. Specifically, it retrieves 50 closest images to each class text based on the CLIP embeddings. They can see that the CLIP embedding sets of the *bird* and *frog* clusters are indeed quite close, and they can find that the highlighted point corresponding to the data item selected in the concept view is at the border of the *bird* cluster. In fact, when users see the image, they find that this item is actually *frog* but misclassified as *bird*. By retrieving similar images in the instance retrieval view, users can further understand that this is because there are some birds and frogs with similar colors and outlines. Subsequently, users perform point-set alignment by dragging this point closer to the correct *frog* cluster, as shown in Fig. 9.

In the backend, we use the ground truth to verify that our visual alignment is indeed helpful. Specifically, before the alignment, the overall accuracy is 69.28%. Particularly, the category with the lowest accuracy is *frog*, with only 32.82%. After the visual alignment of only one single data point in this case, the accuracy for *frog* category rises to 45.24% among the 10,000 images in batch 1 of CIFAR-10. At the same time, overall accuracy also increases to 70.66%.

6.2 Case 2: Instance Retrieval and Compositional Logic

In this case, we show how our system can be used to visualize and refine the compositional logic in instance retrieval. Specifically, we test cross-modal retrieval on the COCO 2017 dataset [44]. The training set of the COCO dataset consists of more than 110,000 images with annotations of captions and objects in the image from 80 categories.

Table 2: Accuracy of *"person and bear"* before and after the alignment.

	top 5	top 20	top 30	top 40	top 50
Before	60.00%	50.00%	43.33%	40.00%	40.00%
After	**80.00%**	**60.00%**	**53.33%**	**52.50%**	**52.00%**

For the retrieval, users enter two queries: the first one is a simple keyword: *"person"* while the second one is composed of two elements using natural language logic expression: *"person and bear"*. Users first leverage our MFM method to project the respective top 50 image results of the two queries in relation to the keyword texts in CLIP embeddings. In the contextual projection shown in Fig. 10 (a), users can see that there is an obvious dense cluster near the text embedding of the composed query *"person and bear"*. In contrast, the distribution of image embeddings is rather sparse near the text embedding of the simple query *"person"*. When users mouse over some data points to explore particular instances, they find that the retrieval results of *"person"* are more diverse than those of *"person and bear"*. More importantly, they even find that the cluster actually contains many similar incorrect results only containing bear in the wild without any person, which shows that the CLIP embeddings do not sufficiently understand the logic in *"person and bear"*. Users can further verify this finding by adding another keyword query *"bear"* and visualize the results together, as shown in Fig. 5 (b).

Upon identifying the misalignment issue for the composed query, users can proceed to interactively align the CLIP embeddings in the system's align mode. Specifically, in the projection view, they first lasso to select samples of the incorrect cluster, which are added to the first alignment subset represented by pink color. Next, they also discover some individual samples of correct images containing both person and bear near the text embedding, which are added to the second alignment subset represented in blue. Subsequently, users can see that the incorrect subset and correct samples are quite close and hard to separate. Finally, users can perform set alignment by dragging the incorrect subset farther away, triggering fine-tuning in the backend.

After the alignment, users can exploit the new CLIP embeddings to re-rank the previous top 500 results. We implement re-ranking instead of completely indexing all the data points in the dataset for faster system reaction. This only takes a few seconds, together with the few-shot alignment. To verify quantitatively that such alignment is indeed helpful, in the backend, we calculate the top k accuracy of the new results compared to previous results, as shown in the table.

6.3 Case 3: Concept Injection and Disentanglement in Cross-modal generation

In this case, we show how our system can be used for alignment in cross-modal generation, with examples of aligning text-to-image Stable Diffusion model. Particularly, compared to specialized AIGC fine-tuning tools such as IntentTuner [76], which only focuses on data augmentation and training functions like LoRA [35] and DreamBooth [56], our visual probing framework allows users to visually inspect and compare the generation results, augmentation data, and prompt keywords before and after fine-tuning through embedding visualization. We choose Stable Diffusion V1-4, which uses CLIP as the input encoder.

The first example (Fig. 11) shows the case of alignment for concept injection, where the pre-trained model does not understand a concept, and users try to inject it into the model's knowledge. For example, as shown in Fig. 11, users may want to input prompt *"Rem rezero"*, which is the name of an animation character, to generate images of the character. However, after the generation of 50 samples by the original model, users can find that our system detects some visual keywords such as *"purple hair"* that is unexpected since the desired character has a prominent feature of short blue hair. Users also enter another concept keyword *"maid"*, which describes the signature dressing style of the expected character. Then, MFM produces a joint projection of the keywords and the generated images, as shown in Fig. 11 (a). Users can find in this projection view many abnormal results. For example, the outliers like Fig. 11 (a) (1) are images of realistic photos. Users can also

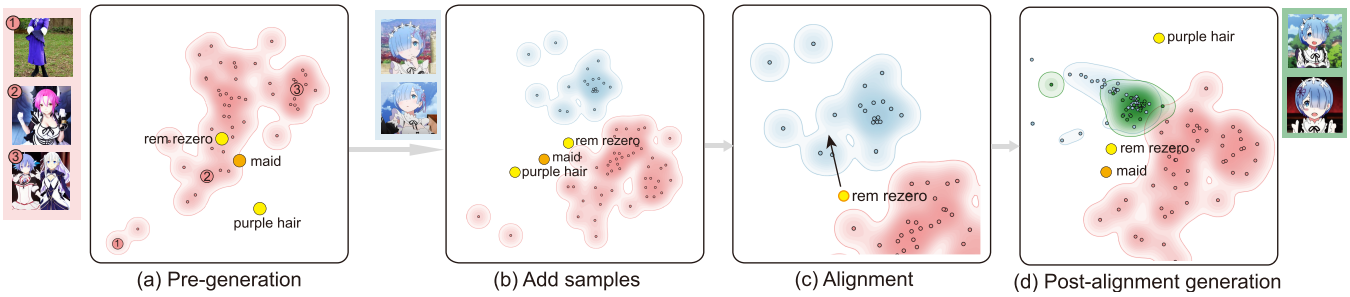


Fig. 11: Alignment for text-to-image cross-modal generation. The first case shows concept injection, where the original model does not understand a concept, and users provide visual samples to align with the textual concept keywords.

confirm that some images close to the *"purple hair"* text embeddings contain distinct purple hair from the expected blue hair (Fig. 11 (a) (2)).

To align the model with this new concept, users collect 20 correct sample images of Rem and upload them to the system. The newly uploaded samples are added to the projection as shown in Fig. 11 (b). We can see the projection view clearly shows the gap between the pre-generated results (red) and the correct samples (blue), while the prompt keyword *"rem rezero"* sits somewhere in the middle between the two clusters, indicating insufficient alignment to the correct set. After observing this, users can directly drag the text embedding point of *"rem rezero"* towards the correct sample set, as shown in Fig. 11 (c). This action triggers the backend alignment process. When the alignment is finished, the system generates 50 samples (green) of *"rem rezero"* with the newly-aligned model (Fig. 11 (d)). We can see that compared to the red set, the green set are much closer to the correct sample cluster, which signifies successful alignment to the correct concept. Users further explore the details of the new generated images and find that they have captured the most important features of the desired character, including the signature short blue hair and the maid dress.

The second example (Fig. 1) showcases alignment for disentangling the generation. Sometimes, misalignment in the cross-modal generation model causes concepts to entangle, leading to unexpected and uncontrollable generation. For instance, when users want to generate landscape paintings, they enter two prompts with the same subject but different artists' names: *"waterlily pond by Van Goph"* and *"waterlily pond by Monet"*. The system first generates 50 samples for each prompt with the same set of random seeds. Apart from textual keywords like *"Monet"* and *"Van Goph"*, the system also detects an unexpected visual concept: *"bridge"*. Then, by exploring our MFM projection view as shown in Fig. 1 (a), users can find that the points generated by *"waterlily pond by Monet"* (red) are concentrated in a dense cluster close to the text embedding of *"bridge"* while the results of *"waterlily pond by Van Goph"* are distributed more sparsely in the projection space. Inspecting the data points, users can discover that the Monet results have a highly similar composition, almost always containing a bridge. This pattern is more evidently shown in the concept axis in Fig. 1 (a), where the high values on the Monet dimension are strongly bundled with high values on the bridge dimension for the red set. In contrast, the Van Goph results are much more diverse with different compositions. This observation indicates that in the Monet prompt, the name of the artist is highly entangled with the visual concept of the bridge, thus significantly reducing the diversity of the generation. To align the model for disentanglement, users first need some fine-tuning data but may find it difficult to collect Monet's paintings of the waterlily pond manually. To address the issue, they can exploit the weighted embedding function provided by our augmentation panel to generate more disentangled samples. Specifically, users can combine the CLIP embeddings of different keywords and phrases in the original prompt through a weighted sum of the embedding vectors to guide the generation of augmented samples. Users can select from these generated augmentation images satisfactory samples that match the prompt *"waterlily pond by Monet"*. They can then add the samples to the projection (green set) in Fig. 1 (b), where users can find an obvious distance between the pre-generated Monet set and the augmentation set. Next, as shown in Fig. 1 (c), users can drag the pre-generated Monet set contour towards the augmentation set in the projection or drag the box representing the pre-generated Monet set in the concept axis, which triggers the backend alignment process of the two sets. Finally, in Fig. 1 (d), the system will generate a new set (purple) by the same prompt of the original red set (*"waterlily pond by Monet"*) with the post-alignment model. Users can see that compared with the red set, the purple set is more aligned to the green set while having more diversity (containing images with and without bridges).

7 Discussion

Speed limitations. Our system and method have two limitations in terms of speed. First, even though the parametric method can scale to large datasets with shorter asymptotic time (as shown in supplementary), for smaller datasets, it is not as fast as some traditional methods like t-SNE. Second, the speed of the alignment fine-tuning depends on the specific implementation for different tasks. For classification and retrieval tasks, the triplet loss-based fine-tuning only takes a few seconds. However, for the generation task, the commonly used LoRA fine-tuning can take a few minutes. To address this, we can take advantage of the latest accelerated fine-tuning methods such as HyperDreamBooth [57].
Scalability to more modalities. In this study, we only test our system and method on embeddings of two modalities. However, some multi-modal embeddings involve more than two modalities. For example, the ImageBind [24] embedding models incorporate six modalities including images, text, audio, depth, thermal, and IMU data. For these data, we can extract multi-modal semantic features as concepts and extend our concept visualization to cover more modalities. For modalities that are difficult to observe visually, such as audio, we can represent their conceptual features using text or images. Our MFM method can also be improved to visualize different pairs of modalities, such as text-audio and image-audio.
Pixel-level alignment. Even though our study enables various set and point level alignments, sometimes these alignments are not enough for fine-grained cross-modal tasks. For example, embedding-based objection detection requires sub-instance pixel level alignment [79]. Our research has not so far touched upon this type of sub-instance alignment. Regarding this problem, in future work, our system can integrate interactive semantic segmentation such as Segment Anything Model [39] into the alignment process to allow users to emphasize certain parts of the image they want the model to understand.

8 Conclusion

In this study, we propose a visual probing and alignment framework for exploring and interactively refining multi-modal embeddings. Particularly, for visual probing, we address the modality gap problem by developing a dimension reduction method called Modal Fusion Map (MFM) to optimize the display of inter-modal embedding features. To facilitate interactive alignment, we design an interaction scheme supporting various alignment intentions including point-set and set-set alignment. As shown in our quantitative evaluation and case studies, our framework can help intuitive visual probing and alignment for diverse tasks. This shows the opportunities for future research to increase human moderation of large models that are growing in size but decreasing in transparency.

Acknowledgments

The authors wish to thank the anonymous reviewers for their valuable comments. This work is supported partially by National Natural Science Foundation of China (62172398), and the Guangzhou Basic and Applied Basic Research Foundation (2024A04J6462, 2023A03J0142).

References

[1] CLIP-Interrogator. https://github.com/pharmapsychotic/clip-interrogator. 6

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic propositional image caption evaluation. In *Proc. ECCV*, pp. 382–398. Springer, 2016. doi: 10.1007/978-3-319-46454-1_24 1

[3] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *Proc. CVPRW*, pp. 4959–4968, 2022. doi: 10.1109/CVPRW56347.2022.00543 1

[4] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proc. CVPR*, pp. 21466–21474, 2022. doi: 10.1109/CVPR52688.2022.02080 2

[5] I. Borg and P. J. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005. 2

[6] A. Bougouin, F. Boudin, and B. Daille. TopicRank: Graph-based topic ranking for keyphrase extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 543–551, 2013. 6

[7] C. Chen, J. Wu, X. Wang, S. Xiang, S.-H. Zhang, Q. Tang, and S. Liu. Towards better caption supervision for object detection. *IEEE Trans. Vis. Comput. Graph.*, 28(4):1941–1954, 2021. doi: 10.1109/TVCG.2021.3138933 3

[8] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. doi: 10.1007/s11633-022-1369-5 3

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pp. 1597–1607. PMLR, 2020. doi: 10.48550/arXiv.2002.05709 2

[10] S. Cheng and K. Mueller. The data context map: Fusing data and attributes into a unified display. *IEEE Trans. Vis. Comput. Graph.*, 22(1):121–130, 2015. doi: 10.1109/TVCG.2015.2467552 2, 4

[11] S. Cheng, W. Xu, and K. Mueller. RadViz deluxe: An attribute-aware display for multivariate data. *Processes*, 5(4):75, 2017. doi: 10.3390/pr5040075 2

[12] D. Choi, B. Drake, and H. Park. Co-embedding multi-type data for information fusion and visual analytics. In *Proceedings of International Conference on Information Fusion (FUSION)*, pp. 1–8, 2023. doi: 10.23919/FUSION52260.2023.10224157 2

[13] J. Choo, S. Bohn, G. C. Nakamura, A. M. White, and H. Park. Heterogeneous data fusion via space alignment using nonmetric multidimensional scaling. In *Proceedings of the International Conference on Data Mining*, pp. 177–188. SIAM, 2012. doi: 10.1137/1.9781611972825.16 2, 4, 6

[14] J. J. Y. Chung and E. Adar. PromptPaint: Steering text-to-image generation through paint medium-like interactions. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17, 2023. doi: 10.1145/3586183.3606777 6

[15] F. L. Dennig, M. Miller, D. A. Keim, and M. El-Assady. Fs/ds: A theoretical framework for the dual analysis of feature space and data space. *IEEE Trans. Vis. Comput. Graph.*, 30(8):5165–5182, 2024. doi: 10.1109/TVCG.2023.3288356 2

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423 2

[17] A. Endert, C. Han, D. Maiti, L. House, and C. North. Observation-level interaction with statistical models for visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 121–130. IEEE, 2011. doi: 10.1109/VAST.2011.6102449 2, 3

[18] D. P. Faith, P. R. Minchin, and L. Belbin. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69:57–68, 1987. doi: 10.1007/BF00038687 5

[19] Y. Feng, X. Wang, K. K. Wong, S. Wang, Y. Lu, M. Zhu, B. Wang, and W. Chen. PromptMagician: Interactive prompt engineering for text-to-image creation. *IEEE Trans. Vis. Comput. Graph.*, 30(1):295–305, 2024. doi: 10.1109/TVCG.2023.3327168 3

[20] T. Fujiwara, X. Wei, J. Zhao, and K.-L. Ma. Interactive dimensionality reduction for comparative analysis. *IEEE Trans. Vis. Comput. Graph.*, 28(1):758–768, 2021. doi: 10.1109/TVCG.2021.3114807 2, 3, 4

[21] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of International Conference on Learning Representations*, 2022. 3

[22] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao. Clip-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x 7

[23] Y. Gao, J. Liu, Z. Xu, T. Wu, E. Zhang, K. Li, J. Yang, W. Liu, and X. Sun. SoftCLIP: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 1860–1868, 2024. doi: 10.1609/aaai.v38i3.27955 1

[24] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *Proc. CVPR*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457 1, 3, 9

[25] A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. In L. Saul, Y. Weiss, and L. Bottou, eds., *Proc. NIPS*, vol. 17. MIT Press, 2004. 2

[26] M. Gönen. Embedding heterogeneous data by preserving multiple kernels. In *Proc. ECAI*, pp. 381–386. IOS Press, 2014. doi: 10.3233/978-1-61499-419-0-381 2

[27] L. Gou, L. Zou, N. Li, M. Hofmann, A. K. Shekar, A. Wendt, and L. Ren. VATLD: A visual analytics system to assess, understand and improve traffic light detection. *IEEE Trans. Vis. Comput. Graph.*, 27(2):261–271, 2020. doi: 10.1109/TVCG.2020.3030350 3

[28] Y. Guo, H. Shao, C. Liu, K. Xu, and X. Yuan. PromptTHis: Visualizing the process and influence of prompt editing during text-to-image creation. *IEEE Trans. Vis. Comput. Graph.*, pp. 1–12, 2024. doi: 10.1109/TVCG.2024.3408255 3

[29] W. He, L. Zou, A. K. Shekar, L. Gou, and L. Ren. Where can we help? a visual analytics approach to diagnosing and improving semantic segmentation of movable objects. *IEEE Trans. Vis. Comput. Graph.*, 28(1):1040–1050, 2021. doi: 10.1109/TVCG.2021.3114855 3

[30] F. Heimerl and M. Gleicher. Interactive analysis of word vector embeddings. *Comput. Graph. Forum*, 37(3):253–265, 2018. doi: 10.1111/cgf.13417 2, 4

[31] F. Heimerl, C. Kralj, T. Möller, and M. Gleicher. EmbComp: Visual interactive comparison of vector embeddings. *IEEE Trans. Vis. Comput. Graph.*, 28(8):2953–2969, 2020. doi: 10.1109/TVCG.2020.3045918 2

[32] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595 1

[33] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. DNA visual and analytic data mining. In *Proceedings of IEEE VIS*, pp. 437–441, 1997. doi: 10.1109/VISUAL.1997.663916 2

[34] M. C. Hout, M. H. Papesh, and S. D. Goldinger. Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):93–103, 2013. doi: 10.1002/wcs.1203 5

[35] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of International Conference on Learning Representations*, 2021. 3, 7, 8

[36] Y. Huang, F. Shakeri, J. Dolz, M. Boudiaf, H. Bahig, and I. Ben Ayed. LP++: A surprisingly strong linear probe for few-shot clip. In *Proc. CVPR*, pp. 23773–23782, 2024. doi: 10.48550/arXiv.2404.02285 7

[37] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*, pp. 4904–4916. PMLR, 2021. doi: 10.48550/arXiv.2102.05918 2

[38] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4(1):1–13, 2003. doi: 10.1186/1471-2105-4-48 6

[39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proc. ICCV*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371 9

[40] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[41] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. doi: 10.1109/ACCESS.2020.3031549 3

[42] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. ICML*, pp. 19730–19742. PMLR, 2023. doi: 10.5555/3618408.3619222 6

[43] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Proc. NIPS*, 35:17612–17625, 2022. doi: 10.5555/3600270.3601550 2, 4, 6

[44] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48 8

[45] S. Liu, P.-T. Bremer, J. J. Thiagarajan, V. Srikumar, B. Wang, Y. Livnat, and V. Pascucci. Visual exploration of semantic relationships in neural word embeddings. *IEEE Trans. Vis. Comput. Graph.*, 24(1):553–562, 2017. doi: 10.1109/TVCG.2017.2745141 2, 4

[46] Y. Liu, E. Jun, Q. Li, and J. Heer. Latent space cartography: Visual analysis of vector space embeddings. *Comput. Graph. Forum*, 38(3):67–78, 2019. doi: 10.1111/cgf.13672 2, 4

[47] J. Lu, D. Batra, D. Parikh, and S. Lee. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Proc. NIPS*, 32, 2019. doi: 10.5555/3454287.3454289 3

[48] M. Meyer, A. Barr, H. Lee, and M. Desbrun. Generalized barycentric coordinates on irregular polygons. *Journal of Graphics Tools*, 7(1):13–22, 2002. doi: 10.1080/10867651.2002.10487551 2

[49] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Proc. NIPS*, 26:1–9, 2013. doi: 10.5555/2999792.2999959 2

[50] Y. Ouali, A. Bulat, B. Matinez, and G. Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *Proc. ICCV*, pp. 15534–15546, 2023. doi: 10.1109/ICCV51070.2023.01424 2, 3, 6

[51] M. Quist and G. Yona. Distributional scaling: An algorithm for structure-preserving embedding of metric and nonmetric spaces. *The Journal of Machine Learning Research*, 5:399–420, 2004. doi: 10.5555/1005332.1005346 5

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pp. 8748–8763, 2021. doi: 10.48550/arXiv.2103.00020 1, 2, 3, 7

[53] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. doi: 10.48550/arXiv.2204.06125 1, 2

[54] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042 3

[55] S. Ruder, I. Vulić, and A. Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019. doi: 10.1613/jair.1.11640 3

[56] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155 8

[57] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proc. CVPR*, pp. 6527–6536, 2024. doi: 10.48550/arXiv.2307.06949 9

[58] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, pp. 815–823, 2015. doi: 10.1109/CVPR.2015.7298682 7

[59] M. Skrodzki, H. van Geffen, N. F. Chaves-de Plaza, T. Höllt, E. Eisemann, and K. Hildebrandt. Accelerating hyperbolic t-sne. *IEEE Trans. Vis. Comput. Graph.*, 30(7):4403–4415, 2024. doi: 10.1109/TVCG.2024.3364841 2

[60] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Trans. Vis. Comput. Graph.*, 29(1):1146–1156, 2022. doi: 10.1109/TVCG.2022.3209479 3

[61] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008. 2

[62] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proc. CVPR*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087 1

[63] Q. Wang, S. L'Yi, and N. Gehlenborg. DRAVA: Aligning human concepts with machine learning latent dimensions for the visual exploration of small multiples. In *Proc. ACM CHI*, pp. 1–15, 2023. doi: 10.1145/3544548.3581127 2, 3

[64] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987. doi: 10.1016/0169-7439(87)80084-9 2

[65] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proc. CVPR*, pp. 1900–1910, 2023. doi: 10.1109/CVPR52729.2023.00189 6

[66] T. Wu, E. Jiang, A. Donsbach, J. Gray, A. Molina, M. Terry, and C. J. Cai. PromptChainer: Chaining large language model prompts through visual programming. In *CHI Extended Abstracts*, pp. 1–10, 2022. doi: 10.1145/3491101.3519729 3

[67] Y. Wu, T. Mu, P. Liatsis, and J. Y. Goulermas. Computation of heterogeneous object co-embeddings from relational measurements. *Pattern Recognition*, 65:146–163, 2017. doi: 10.1016/j.patcog.2016.12.004 2

[68] J. Xia, L. Huang, W. Lin, X. Zhao, J. Wu, Y. Chen, Y. Zhao, and W. Chen. Interactive visual cluster analysis by contrastive dimensionality reduction. *IEEE Trans. Vis. Comput. Graph.*, 29(1):734–744, 2022. doi: 10.1109/TVCG.2022.3209423 2, 3, 4, 5, 6

[69] J. Xia, L. Huang, Y. Sun, Z. Deng, X. L. Zhang, and M. Zhu. A parallel framework for streaming dimensionality reduction. *IEEE Trans. Vis. Comput. Graph.*, 30(1):142–152, 2023. doi: 10.1109/TVCG.2023.3326515 5

[70] X. Xie, X. Cai, J. Zhou, N. Cao, and Y. Wu. A semantic-based method for visualizing large image collections. *IEEE Trans. Vis. Comput. Graph.*, 25(7):2362–2377, 2018. doi: 10.1109/TVCG.2018.2835485 2

[71] W. Yang, M. Liu, Z. Wang, and S. Liu. Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media*, pp. 1–26, 2024. doi: 10.1007/s41095-023-0393-x 3

[72] Y. Ye, J. Hao, Y. Hou, Z. Wang, S. Xiao, Y. Luo, and W. Zeng. Generative AI for visualization: State of the art and future directions. *Visual Informatics*, 2024. doi: 10.1016/j.visinf.2024.04.003 1

[73] Y. Ye, R. Huang, and W. Zeng. VISAtlas: An image-based exploration and query system for large visualization collections via neural image embedding. *IEEE Trans. Vis. Comput. Graph.*, 2022. doi: 10.1109/TVCG.2022.3229023 2

[74] Y. Ye, Q. Zhu, S. Xiao, K. Zhang, and W. Zeng. The contemporary art of image search: Iterative user intent expansion via vision-language model. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):Article 180:1–31, 2024. doi: 10.1145/3641019 1

[75] Z. Zang, S. Cheng, L. Lu, H. Xia, L. Li, Y. Sun, Y. Xu, L. Shang, B. Sun, and S. Z. Li. DMT-EV: An explainable deep network for dimension reduction. *IEEE Trans. Vis. Comput. Graph.*, 30(3):1710–1727, 2022. doi: 10.1109/TVCG.2022.3223399 5

[76] X. Zeng, Z. Gao, Y. Ye, and W. Zeng. IntentTuner: An interactive framework for integrating human intents in fine-tuning text-to-image generative models. In *Proc. ACM CHI*, pp. 182:1–18, 2024. doi: 10.1145/3613904.3642165 8

[77] X. Zeng, H. Lin, Y. Ye, and W. Zeng. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *arXiv preprint arXiv:2407.20174*, 2024. doi: 10.48550/arXiv.2407.20174 2

[78] X. Zhang, S. Cheng, and K. Mueller. Graphical enhancements for effective exemplar identification in contextual data visualizations. *IEEE Trans. Vis. Comput. Graph.*, 29(9):3775–3787, 2022. doi: 10.1109/TVCG.2022.3170531 2, 5

[79] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, et al. RegionCLIP: Region-based language-image pretraining. In *Proc. CVPR*, pp. 16793–16803, 2022. doi: 10.1109/CVPR52688.2022.01629 9

[80] C. Zhou, F. Zhong, and C. Öztireli. CLIP-PAE: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–9, 2023. doi: 10.1145/3588432.3591532 2, 4, 6