






DKMap: Interactive Exploration of Vision-Language Alignment in Multimodal Embeddings via Dynamic Kernel Enhanced Projection

Yilin Ye , Chenxi Ruan , Yu Zhang , Zikun Deng , and Wei Zeng 

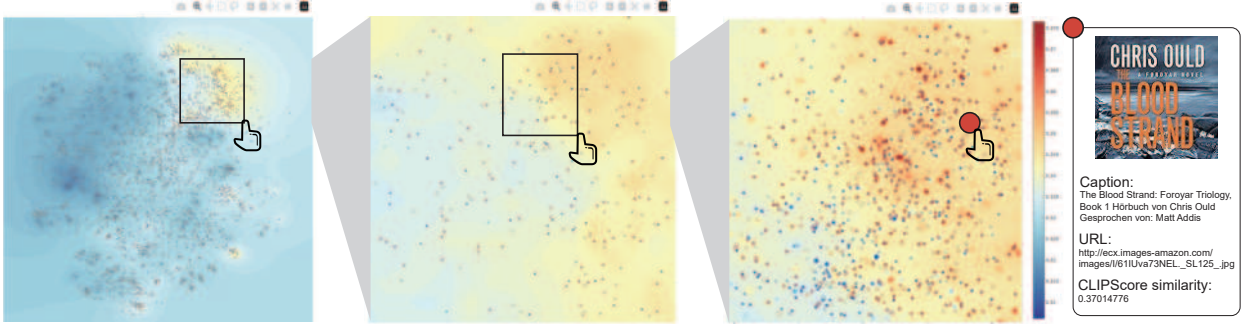


Fig. 1: **Multi-scale exploration of LAION-400M [46] CLIP embeddings using DKMap.** A Plotly-based implementation of *DKMap* for computational notebook features dynamic zooming, which can progressively reveal local details with kernel enhanced projection, enabling the exploration of instance-level attributes.

Abstract— Examining vision-language alignment in multimodal embeddings is crucial for various tasks, such as evaluating generative models and filtering pretraining data. The intricate nature of high-dimensional features necessitates dimensionality reduction (DR) methods to explore alignment of multimodal embeddings. However, existing DR methods fail to account for cross-modal alignment metrics, resulting in severe occlusion of points with divergent metrics clustered together, inaccurate contour maps from over-aggregation, and insufficient support for multi-scale exploration. To address these problems, this paper introduces *DKMap*, a novel DR visualization technique for interactive exploration of multimodal embeddings through **D**ynamic **K**ernel enhanced projection. First, rather than performing dimensionality reduction and contour estimation sequentially, we introduce a *kernel regression supervised t-SNE* that directly integrates post-projection contour mapping into the projection learning process, ensuring cross-modal alignment mapping accuracy. Second, to enable multi-scale exploration with dynamic zooming and progressively enhanced local detail, we integrate *validation-constrained α refinement* of a generalized t-kernel with *quad-tree-based* multi-resolution technique, ensuring reliable kernel parameter tuning without overfitting. *DKMap* is implemented as a multi-platform visualization tool, featuring a web-based system for interactive exploration and a Python package for computational notebook analysis. Quantitative comparisons with baseline DR techniques demonstrate *DKMap*'s superiority in accurately mapping cross-modal alignment metrics. We further demonstrate generalizability and scalability of *DKMap* with three usage scenarios, including visualizing million-scale text-to-image corpus, comparatively evaluating generative models, and exploring a billion-scale pretraining dataset.

Index Terms—Kernel Regression, Vision-language Alignment, Multimodal Embeddings, Interactive Exploration

1 INTRODUCTION

Multimodal models, including vision-language models (VLMs) [43], text-to-image (T2I) models [45], and multimodal large language models (MLLM) [33], are widely used for applications such as text and image data retrieval and generation. These models rely on multimodal embeddings, which align representations of different modalities in high-dimensional spaces. For example, pretrained multimodal embeddings, e.g., CLIP [43], can be directly applied to tasks such as zero-shot classification and cross-modal retrieval (e.g., [72]). The embeddings are also commonly used to measure alignment between input and output in cross-modal generative models, such as text-to-image and text-to-

3D models [30, 45]. Specifically, in such scenarios, vision-language alignment is an important property which refers to how well the visual modality matches the text modality. When evaluating the alignment by multimodal embeddings, studies typically employ end-to-end methods and report metrics, such as CLIPScore [21] or human preference [60]. These metrics can all be termed as cross-modal alignment metrics, which typically rely on operations in multimodal embedding space such as dot product to quantify vision-language alignment. However, such averaged metrics are insufficient for capturing alignment distribution and identifying extreme values. An alternative approach is to use dimensionality reduction (DR) methods to project high-dimensional embeddings into 2D space and color-code points based on cross-modal alignment metrics, such as CLIPScore, as illustrated in Fig. 2(a).

However, traditional DR methods, such as PCA [1], t-SNE [51], and UMAP [38], seek to preserve properties such as variance and neighborhood distribution in high-dimensional features, but are not tailored for accurate large-scale dataset overview and exploration (especially alignment metrics distribution). Without considering alignment metrics, the resulting scatterplots often suffer from serious occlusion, where points with highly divergent metrics are clustered closely together, making it difficult to discern the metric distribution, as highlighted in Fig. 2(a-1&2). Moreover, such divergent metric distributions can lead to over-averaging in continuous contour estimation, which is commonly

- Yilin Ye, Chenxi Ruan, and Wei Zeng (corresponding author) are with the Hong Kong University of Science and Technology (Guangzhou). Yilin Ye and Wei Zeng are also with the Hong Kong University of Science and Technology. E-mail: {yyebd@connect., weizeng@hkust-gz.edu.cn.
- Yu Zhang is with University of Oxford. E-mail: yuzhang94@outlook.com.
- Zikun Deng is with South China University of Technology. E-mail: zkdeng@scut.edu.cn.

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.xxx/TVCG.201x.xxxxxxx

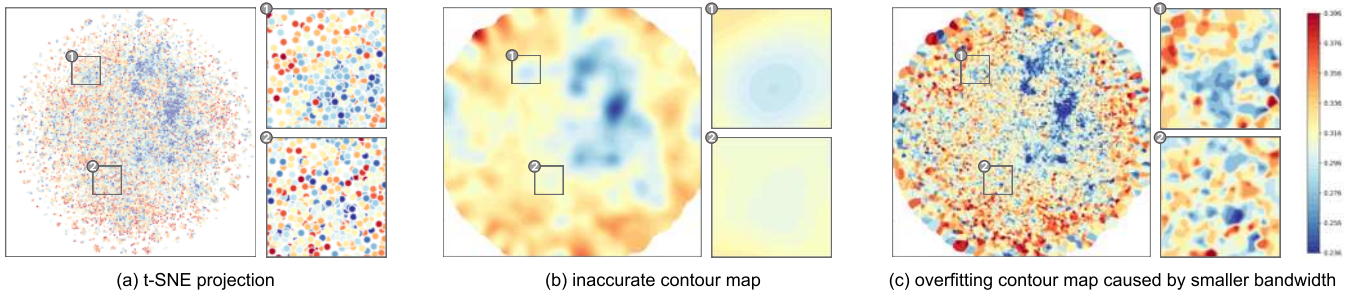


Fig. 2: **Limitations of traditional DR methods when visualizing multimodal embeddings:** (a) Scatterplots by t-SNE do not clearly reveal metric distribution due to local occlusion and dense neighborhoods with highly diverging metric values. (b) Contour mapping based on the projected scatterplot suffers from inaccurate mapping due to over-averaging effect by auto-selected bandwidth. (c) Smaller bandwidth for the contour map causes overfitting with spurious contours.

used for large-scale embeddings visualization; see Fig. 2(b). Recently, visualization and machine learning researchers have proposed specific visualization methods for multimodal embeddings [3, 66, 71]. However, these methods, such as ModalChorus [71], treat text and image embeddings separately, only using the text embeddings as anchored to enhance contextual layout of image embeddings, without a unified point representation for text-image pairs. This deficiency makes it difficult to explicitly encode vision-language alignment in the projection space. In addition, existing supervised methods for general DR problems mostly focus on discrete categorical supervision [47, 54] or suffer from limited scalability, lack of explicit supervision in visualization space, and insufficient multi-scale interaction [6, 47, 59]. These factors prohibit their application to large-scale multimodal embeddings with continuous alignment metric.

To improve the visualization of vision-language alignment in multimodal embeddings, we identify two key challenges to be addressed. The first challenge lies in accurate overview for large multimodal datasets. It involves faithfully representing alignment metric distributions while bridging the gap between high- and low-dimensional spaces. Since these alignment metrics are computed from high-dimensional embeddings, the DR process necessarily introduces distortions and information loss that obscure the original relationships. This increases the difficulty of predicting the function value distribution on the projection space with 2D-sample-based kernel contour mapping. Second, it is challenging to increase detail of the visualization interactively on user zooming. For example, it is difficult to adjust parameters (e.g., bandwidth) in traditional kernel contour mapping to show more local detail on user demand, as the commonly used Gaussian RBF kernel is highly sensitive to bandwidth. Directly adjusting the bandwidth can easily cause overfitting with even larger mapping error and abrupt change of the landscape, as shown in Fig. 2(c).

To fill the gap, we propose *DKMap*, a novel mapping method for multi-scale interactive visualization of multimodal embeddings. The core of *DKMap* is a dynamic kernel that can enhance the accuracy of post-projection contour mapping and smoothly increase local detail on user zooming. Specifically, we first develop a supervised parametric t-SNE method explicitly guided by post-projection kernel regression loss. To bridge between parametric projection and non-parametric kernel regression, we jointly learn the parameters of a generalized t-kernel with a projection network through a weighted train-validation MSE loss. Then, to support multi-scale exploration with dynamic zooming, we propose a view-dependent dynamic contour map technique combining multi-resolution with kernel refinement, which mitigates the risks of overfitting and abrupt change. To support the application of *DKMap*, we develop a web-based interactive visualization system to allow easily accessible browser exploration. We also implement a Python package that can be seamlessly integrated into interactive computational notebook for data science experts usage. To evaluate *DKMap*, we perform quantitative experiment on mapping accuracy along with a user study to show the advantages of *DKMap* compared to traditional methods. We also demonstrate the utility of our methods in usage scenarios, including visualizing generated data, comparing generative models, and visualizing enormous-scale pretraining dataset. Demo and source code of *DKMap*

are available at: <https://github.com/HKUST-CIVAL/DKMap>.

The contributions of our study are as follows:

- A novel dynamic kernel enhanced dimension reduction method *DKMap* that can simultaneously learn the projection of high-dimensional points and post-projection contour mapping via kernel regression supervised parametric t-SNE. The dynamic kernel can further support detail enhanced zooming by combining refinement of generalized t-kernel with multi-resolution technique to dynamically adjust granularity of contour map while mitigating the risks of overfitting and abrupt change.
- Multi-platform visualization tools, including a web-based visualization system and a Python package for accessible and flexible data exploration on browsers and computational notebook.
- Extensive experiments, including quantitative evaluation on accuracy of metric mapping, user study for human evaluation of the mapping, and three usage scenarios covering million-scale generated corpus visualization, evaluation of specific generative models, and billion-scale pretraining dataset visualization.

2 RELATED WORK

2.1 Vision-Language Alignment

Vision-language alignment plays a crucial role in cross-modal learning, with multimodal embeddings serving as its foundation. For example, CLIPScore [21] measured in Contrastive Language-Image Pre-Training (CLIP) [43] space provides a metric for measuring the alignment between text prompts and generated images. Recent work has introduced human preference metrics, including HPS [60, 61], PickScore [24], and ImageReward [65], to better align with human judgments. However, these metrics only provide an aggregated score. The black-box nature limits interpretability, preventing users from examining the underlying score distribution. This limitation underscores the need for embedding visualization techniques that can reveal fine-grained value distributions across the cross-modal embedding space. Specifically, in this study we are concerned with multimodal embeddings' capability to measure vision-language alignment through pairwise embedding distance in the high-dimensional space. This has a wide range of applications in training [44], fine-tuning [65] and evaluation [60] of generative models such as text-to-image [45], image editing [76] and text-to-3D models [30], and filtering [46] and retrieval [72] of large-scale data.

Many studies [31, 55, 76] in machine learning have employed traditional DR techniques such as PCA [1], t-SNE [51], and UMAP [38] to visualize vision-language alignment. However, these methods process each modality independently, causing modality gap and obscuring the underlying distribution of alignment metrics. As a result, they struggle to differentiate well-aligned from misaligned samples in the joint embedding space, resulting in points with divergent alignment scores clustered together (see Fig. 2(a)). The limitation further leads to unreliable contour mapping, where the over-averaging effect of mixed metrics introduces inaccuracies (see Fig. 2(b)). Moreover, the inability to incorporate additional metrics restricts comparative analysis to side-by-side visualizations of model performance. For example, Embedding

Comparator [2] arranges two t-SNE projections adjacent to each other for model comparison, which fails to establish sample-level alignment relationships between the projections.

While there exist supervised DR approaches that can integrate target variable into the projection, most of them are limited to categorical classification labels [54, 62], which inadequately account for continuous metric distributions. A few supervised methods that can incorporate continuous metric such as St-SNE [6] and Fisher t-SNE [47] suffer from severe limitations in scalability to typically million-scale multimodal datasets. They also lack explicit supervision from visualization space to facilitate accurate overview and multi-scale exploration. In addition, neural-network-based projection methods such as autoencoders [11, 12] can project larger scale datasets but suffer from difficult and unstable training (requiring long epochs and producing inaccurate landscape), especially for complex multimodal embeddings. This study fills the gap by introducing a novel kernel regression-based supervised DR that simultaneously learns point projections and contour mapping of alignment metrics. The advantages of our method are prominent, including more accurate representation of alignment metrics, support for multi-scale exploration of large-scale datasets, and explicit encoding of metric differences.

2.2 Embedding-Based Visual Exploration

High-dimensional embeddings encode raw data into dense vector representations, while preserving their semantic and structural relationships. Depending on the number of modalities involved, these embeddings can be categorized as *unimodal* or *multimodal*. Unimodal embeddings serve as the foundation for various machine learning tasks, including classification, retrieval, and generation. To facilitate the understanding and analysis of these representations, numerous visual exploration systems have been developed, supporting tasks such as data filtering [35, 70, 74], embedding comparison [2, 20, 29], and model interpretation [28, 52, 53]. These systems share some common features: leveraging DR methods to project high-dimensional embeddings into lower-dimensional spaces, incorporating semantic axes to reveal conceptual relationships within the projected space, and integrating attribute views as complementary visual representations of feature spaces. For example, Latent Space Cartography [35] integrates DR methods of t-SNE, UMAP, and PCA, enables users to interactively explore semantic relationships in latent space through semantic axes, and provide complementary views for attribute vectors.

The emergence of multimodal models such as CLIP [43] and ALIGN [22] creates an increasing demand for effective visualization of multimodal embeddings. However, the task presents unique challenges due to the heterogeneous distributions of different modalities. Traditional DR techniques and axis-based interactions designed for unimodal embeddings often fail to adequately represent the complex semantic relationships in multimodal spaces. While context-aware DR techniques [5, 69] can incorporate context information from one modality to guide projections of another, they typically process modalities separately rather than through joint projection. Recent advances have introduced anchoring-based approaches [3, 71] that jointly project embeddings from two modalities into a unified space. For example, ModalChorus [71] relies on text embeddings as anchor points to align and probe corresponding image embeddings. However, these methods face challenges with increasing numbers of anchor points, and lack explicit representation of cross-modal alignment metrics.

This work proposes a supervised DR method that extends parametric t-SNE [17] to support effective visualization of multimodal embeddings and their alignment metrics. Specifically, we adopt a post-projection kernel regression loss with neighborhood constraint to preserve cross-modal correspondence. The supervised approach can be flexibly adapted to optimize the kernels used in contour map generation, supporting large-scale multimodal datasets with millions of data points. For practical deployment, we provide two implementations, including an interactive web-based visualization system and a Python package for computational notebooks.



Fig. 3: **Example of ambiguous neighborhood:** Concatenating the text and image embeddings to form the multimodal embeddings will cause ambiguous neighborhood, with two types of neighboring points for Point 1: (a) points with the same text and some shift in the image part and (b) points with translation in both text and image parts, both with similar distances to Point 1 but larger difference in alignment scores.

2.3 Contour Estimation

Kernel estimation is a commonly used non-parametric statistical technique to approximate the distribution of values such as density [57] and predictive variables [9], with Gaussian kernel [48] most commonly used for smooth estimation. Specifically, in visualization domain, two-dimensional Gaussian kernel estimation map has been widely used in visual analytics applications for displaying the global distribution of density [4] or other performance metrics in machine learning [52]. It is also frequently used in point-map overlay design as an effective graphical enhancement of scatterplot [37, 69].

These applications of Gaussian kernel estimation are also common in contour estimation for DR scatterplots [5, 40, 52, 55, 75]. For example, Data Context Map [5] leverages Gaussian map to visualize density and attribute value distribution after projection of multidimensional data. WizMap [55] utilizes Gaussian contour map to overview large-scale embeddings datasets. However, applying kernel estimation directly to projected high-dimensional data can be misleading. After projection, the local structure of the data, such as density, may significantly differ from the original high-dimensional space [39]. This discrepancy can lead to substantial errors in estimation, regardless of the bandwidth selection method used. In addition, post-projection contour map using Gaussian kernels tend to lack local details due to over-averaging effect. This makes it difficult to increase the detail of the map dynamically as users zoom in to local region. Studies [11, 12] have investigated the usage of auto-encoder-based approaches to estimate continuous contour, which seeks to reconstruct high-dimensional vectors from 2D points. However, for complex multimodal embeddings, the reconstruction is still difficult, potentially leading to large errors in contour mapping.

To resolve this challenge, we integrate kernel estimation with the DR process through supervised projection. This coupling enables dynamic adjustment of the kernels to optimize metric landscape visualization. Moreover, our approach supports dynamic kernels that can adapt to varying levels of detail, facilitating interactive zooming capabilities.

3 DESIGN CONSIDERATION

Multimodal embeddings exhibit fundamental characteristics: high-dimensional vectors, heterogeneous cross-modal distributions, and large-scale data volumes. Based on existing literature in machine learning [55, 76] and visualization [3, 71], along with our experience with vision-language models, we identify the core requirements for effective visualization of vision-language alignment in multimodal embeddings: extending DR method to support multimodal embeddings, and enable interactive multi-scale exploration of the projections. Specifically, we formulate the key design considerations of our work, as follows:

DC1: Projection of concatenated vision-language embeddings.

To visually explore and analyze multimodal embeddings, particularly the vision-language alignment, we need to develop a customized dimensionality reduction method that tackles a special type of input: the concatenated vision-language embeddings $x = (t, v)$ where v and t are the image embedding vector and text embedding vector respectively. The concatenation is necessary because we need to treat a pair of text and image as a single data item when visualizing alignment. The choice of the simple concatenation instead of other feature transformation is

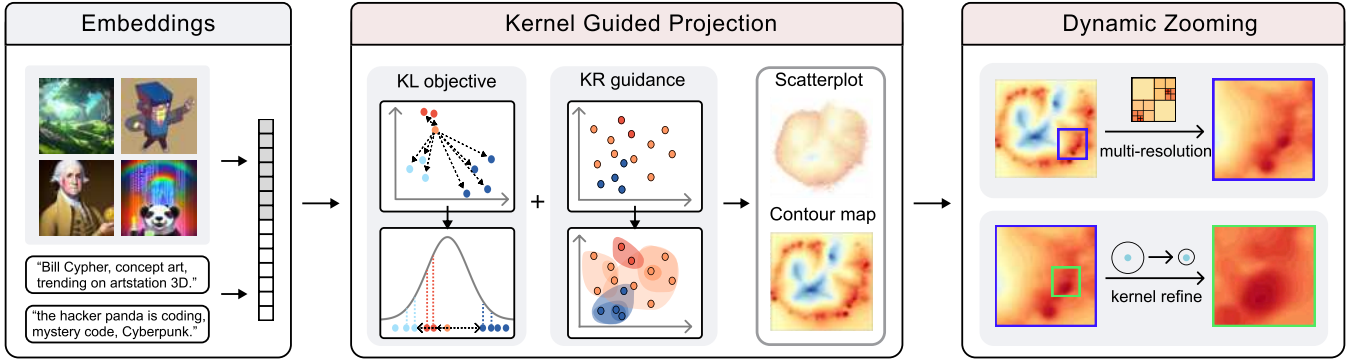


Fig. 4: **DKMap** is a kernel regression-supervised DR method designed to address the challenges of visualizing cross-modal embeddings, with a two-stage approach: 1) Kernel regression-guided projection, which incorporates both the KL objective and kernel regression guidance to enhance projection and contour estimation, thereby more accurately revealing the cross-modal metric distribution. 2) Dynamic zooming, which integrates multi-resolution exploration with kernel refinement to progressively reveal details.

for two reasons. First, this operation does not induce loss of information about the within-pair alignment (e.g., the cosine distance between t and v in alignment metrics such as CLIPScore). Second, it also maintains inter-pair distances (the cosine distance between different pairs of concatenated embeddings are proportional to the sum of distances in the text and image parts).

DC2: Accurate overview. As multimodal vision-language datasets are typically large scale with at least hundreds of thousands points and up to millions of points, the most important consideration when visualizing the alignment is to enable accurate overview of the distribution of alignment metric. Specifically, two types of visualization are commonly used for the overview: the raw discrete scatterplot and the continuous contour map. The DR scatterplot can reveal all data instances but may suffer from severe visual clutter on large datasets, while the contour map can better reveal global distribution by omitting some details.

DC3: Explicit supervision. However, both types of visualizations in **DC2** face challenges in clearly revealing the alignment metric distribution for multimodal embeddings, as shown in our previous example in Fig. 2. Such phenomenon is especially evident in multimodal embeddings due to its unique property. Specifically, as shown in Fig. 3, the concatenated vision-language embeddings will lead to ambiguous neighborhood with two types of close points: (a) one type with shift in only single modality such as the image part (e.g., image generated by a different model for the same prompt) and (b) the other type with translation in both text and image parts (e.g., image generated by the same model with a slightly different prompt). Meanwhile, most alignment metrics such as CLIPScore and HPSv2 rely on cosine distance between the text and image part. As such, translating two vectors (the text embedding and the image embedding) on similar directions (b) will cause smaller change in the cosine distance between than shifting on the image embedding alone (a), since (b) is close to rigid transformation. This means that the multimodal embedding space will be more likely to have neighborhoods with higher divergence in alignment metrics. Such noises in the high-dimensional space will propagate to the projected space by commonly used dimensionality reduction methods such as t-SNE. To reduce the noise and achieve more accurate overview in the visualization space, we need to develop a new supervised DR method pushing the projection to more clearly reveal the metric distribution. In addition, such supervision should be explicitly guided by the mapping in visualization space to enhance overview in **DC2**.

DC4: Multi-scale zooming. Modern multimodal pretraining and generation tasks commonly involve datasets at million (e.g., DiffusionDB [56]) to even hundreds of millions (e.g., LAION-400M [46]) scales, where static visualizations cannot simultaneously reveal both global patterns and local distribution details. Local regions often contain complex patterns only visible when zoomed in, whilst traditional static contour maps fail to meet this multi-scale exploration need. Inspired by view-dependent density mapping techniques [26], we propose a dynamic alignment metric visualization that adaptively increases local

detail during zoom interactions while maintaining visual continuity.

DC5: Embedding comparison. Multimodal embeddings are commonly used to evaluate different generative models [21, 60]. However, existing approaches that visualize different models in separate maps and rely on side-by-side comparison [2, 20] prove inadequate for benchmark analysis; see Fig. 9(a). This limitation stems from the cognitive difficulty of correlating numerous corresponding points across distinct visualizations [7]. To address this, we aim to design a unified comparison map that directly visualizes alignment metric differences between models using explicit encoding [18]. Specifically, in this comparison scenario, the input of our DR is a concatenation of the triplet $x = (t, v_1, v_2)$ where v_1 and v_2 are embeddings of images generated by two different models.

DC6: Multi-platform tools. We aim to support multi-platform deployment to serve diverse user needs, including 1) a responsive web visualization system enabling interactive exploration of large-scale embeddings with filtering, brushing, and dynamic zooming capabilities; and 2) a Python package for embedding projection and analysis in computational notebook. To ensure practical utility, we optimize for computational efficiency to maintain interactive performance with million-scale datasets.

4 DKMap METHOD

In response to the design considerations, in this section we first present **DKMap**, a kernel regression supervised t-SNE designed to address the problem of projecting concatenated vision-language embeddings (**DC1, DC5**). As illustrated in Fig. 4, **DKMap** incorporates a two-stage approach: 1) *kernel guided projection* (Sec. 4.1): Jointly optimizes projection and contour estimation to reduce distortion and better reveal cross-modal metric distribution [68], and 2) *dynamic zooming* (Sec. 4.2): Enables multi-resolution exploration via kernel refinement for adaptive detail inspection.

4.1 Kernel Guided Projection

Commonly used DR techniques (e.g., t-SNE) are well-suited for classification tasks due to their ability to form distinct visual clusters. However, evaluating multimodal models on complex data, where individual instances often contain multiple overlapping concepts, requires moving beyond classification objectives. Inspired by supervised DR approaches [49], we propose to integrate metric distribution information into the projection through supervised learning. Yet, existing supervised methods primarily focus on discrete classification labels, lacking support for continuous metric supervision. To address this, **DKMap** introduces *kernel-guided t-SNE* as its first component, which incorporates a kernel regression-based loss function to estimate and preserve metric distributions during projection for accurate overview (**DC2**). The reasons for choosing t-SNE over other DRs as the backbone to build upon are mainly three folds: first it can deal with complex nonlinear structure in multimodal embeddings; second its parametric version has good scalability; third the kernel formulation of the original t-SNE

can be easier to integrate with post projection contour mapping which also relies on kernel.

KL objective. Our method is a modification of traditional t-SNE method. The basic t-SNE method first estimates conditional probability distributions in point neighborhoods in high-dimensional (P) and low-dimensional (Q) space respectively through pairwise similarities. Then it seeks to minimize the difference between the two distributions measured by KL-divergence:

$$C = KL(P||Q) = \sum_i \sum_{j \neq i}^n p_{ij} \ln \frac{p_{ij}}{q_{ij}}, \quad (1)$$

where p_{ij} denotes the conditional probability that point x_i would pick x_j as its neighbor in P , and q_{ij} for that in Q . Specifically, the high-dimensional distribution is modeled by Gaussian probability density function while the low dimensional distribution is modeled by the Student t-distribution with a single degree of freedom.

$$p_{ij} = \frac{p_{ji} + p_{ilj}}{2n}, \quad p_{ji} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}, \quad (2)$$

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}}. \quad (3)$$

However, to enable flexible adaptation to multi-scale datasets, slightly different from traditional t-SNE that requires manual settings of perplexity to determine the σ_i parameters, we adopt a multi-scale Gaussian similarity approach [8] as detailed in supplemental.

Kernel regression guidance. To incorporate the metric distribution information, it is important to design a proper supervision signal for continuous contour mapping. Particularly, to avoid the noisy neighborhood with points such as those in Fig. 3 causing inaccurate overview as shown in Fig. 2, it is important to introduce explicit guidance in the projected space to push apart the highly divergent points and adapt the kernel-based contour map accordingly (DC3). To achieve this, we introduce a kernel regression guided supervision method. Specifically, we estimate the distribution of metric m in the projected space via kernel regression:

$$\hat{m}(y) = \frac{\sum_{y_k \in Y_t} K_t(y, y_k) \cdot m_k}{\sum_{y_k \in Y_t} K_t(y, y_k)}, \quad (4)$$

$$K_t(y, y_k) = \left(1 + \alpha \|y - y_k\|^{2\beta}\right)^{-1}, \quad (5)$$

where y_k belongs to the training set Y_t . For kernel-based continuous metric distribution estimation, Gaussian kernel is the most frequently used kernel. This is because of its smooth decay along the radius compared to other kernels such as Triangular kernel and Epanechnikov kernel which have shorter or unsmooth support [57]. However, when combined with the t-SNE projection, Gaussian kernel is not optimal due to its sensitivity as we analyze in Eq. (8) in Sec. 4.2. We observe in experiment that such sensitivity often leads to numerical instability in projection learning and overfitting when adjusting bandwidth. Therefore, we adopt a t-distribution-like kernel instead of the most commonly used Gaussian kernel to ensure it is more stable in joint optimization with the projection and more compatible with the t-distribution modeling of low-dimensional neighborhood in t-SNE. In addition, we set the learnable α and β parameters to increase the flexibility of the kernel and account for the gap between high-dimensional and low-dimensional local radius [39].

However, traditional kernel methods are nonparametric and do not have an explicit loss function during the estimation. In addition, if we only use the training set to construct the loss function for kernel estimation, it will compromise the generalizability of nonparametric

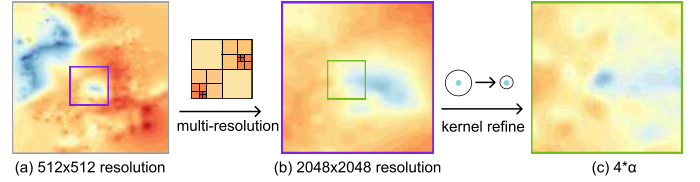


Fig. 5: **Our dynamic zooming approach:** (a) and (b) Multi-resolution zooming for higher level zoom. (c) Kernel refinement with α set to 4 for lower level zoom.

methods and easily lead to overfitting. To address the problem and formulate a proper loss function for metric distribution estimation, we adopt a cross-validation strategy commonly used in kernel methods [48], where we split the projected dataset into training (tr) and validation (vl) subsets Y_{tr} and Y_{vl} of size 9 : 1. Then we construct the loss function using the weighted mean square error on the training set and validation set to balance the within-sample and out-of-sample accuracy:

$$MSE_p = \frac{1}{|Y_p|} \sum_{y \in Y_p} (\hat{m}(y) - m(y))^2, \quad p \in \{tr, vl\}, \quad (6)$$

where $\hat{m}(y)$ is only estimated based on sample points from the training set as in Eq. (3).

Parametric projection objective. Consequently, the overall loss function of $DKMap$ is the sum of the two terms:

$$L = KL(P||Q) + w_1 MSE(m). \quad (7)$$

Finally, to accommodate new incoming data in finetuning, we need to ensure the out-of-sample ability of $DKMap$. In addition, for large datasets, computing $\hat{m}(y)$ with all the training points for each step of gradient descent is expensive. Therefore, we adopt a parametric method that learns the projection via an MLP neural network and divides the whole training dataset into batches for more efficient learning.

4.2 Dynamic Zooming

Next, $DKMap$ leverages a two-stage approach with multi-resolution and kernel refinement to enable dynamic zooming in contour map (DC4). Figure 5 illustrates the effects of these methods.

Multi-resolution technique. To enable fast interactive zooming with dynamic display of contour map detail, we combine our projection result with 2D space partitioning. Specifically, we perform a recursive partitioning of 2D space with a quad-tree. Moreover, for the zooming interaction, traditional KDE cannot dynamically add details of local regions without overfitting. $DKMap$ provides a solution to this problem as it can accurately estimate local detail without the need to change kernel bandwidth. Specifically, the quad-tree constructs a hierarchical division of the 2D projection space into $2^{g_k} \times 2^{g_k}$ number of grids, where g_k is dependent on zoom level with g_0 for the coarsest overview level. Then, for a given zoom level l , we leverage the kernel regression function learned from the parametric t-SNE projection process to estimate the color value $\hat{m}(y_{ij}^{g_k})$ for each grid.

Kernel refinement technique. Although the multi-resolution quad-tree combined with the learned kernel in the parametric t-SNE can already increase the detail when zooming as we describe above, we seek to further enhance the local detail by a view-dependent kernel adjustment technique. Specifically, starting from the first zoom-level of $2^{g_0} \times 2^{g_0}$ grid, we compute a view-dependent refinement of our kernel parameters α . Some previous studies propose dynamically changing the bandwidth of a Gaussian kernel to achieve view-dependent KDE visualization [26]. Here we illustrate why such method can be problematic for visualizing multimodal embeddings, and how our method can alleviate the problem. First, changing the traditional Gaussian bandwidth simply for visual effect without the validation constraint as in equation (Eq. (6)) breaks the statistical validity of kernel methods. This approach risks being a visual trick that misleads users by serious overfitting. Specifically, it tricks users into seeing ‘‘local details’’ which

tend to be a false representation of the true KDE. Second, we perform a sensitivity analysis of the bandwidth in Gaussian kernel compared with the α parameter in our generalized t-kernel to show that changing bandwidth will make the kernel highly sensitive and cause abrupt change, while refining α can achieve more gradual change. Specifically, we denote $K_t = (1 + \alpha \|y\|^{2\beta})^{-1}$ and $K_G = \exp(-\frac{\|y\|^2}{2h^2})$. Then, we take the derivative *w.r.t* h and α respectively:

$$\frac{\partial K_G}{\partial h} = K_G \cdot \frac{\|y\|^2}{h^3}, \quad \frac{\partial K_t}{\partial \alpha} = -\|y\|^{2\beta} K_t^2. \quad (8)$$

Equation Eq. (8) indicates that for small locality corresponding to small $\|y\| < 1$ (since we normalize the projection space), K_G is much more sensitive to h than K_t to α because of the denominator h^3 . In fact, we can achieve a near linear adaptation with α , and keeping β fixed can avoid abrupt exponential change.

Combination of multi-resolution and kernel refinement. The findings above support our adoption of scale refinement in generalized t-kernel for view-dependent contour. However, we avoid directly manipulating α because similar to direct manipulation of bandwidth in Gaussian, such practice disregards statistical validity of the kernel methods. Instead, we incorporate a hybrid approach by combining multi-resolution technique with cross-validation constrained α scaling through a dynamic approach. Specifically, we first use a validation-constraint method for α refinement that avoids compromising the kernel’s statistical performance while achieving zoom-dependent level of local granularity: $\alpha_{k+1} = \alpha_k + \Delta\alpha_k$. The termination of this α updating is subject to the validation set MSE loss constraint in equation (Eq. (6)): $MSE_{v,l}(\alpha_k) \leq MSE_{v,l}(\alpha_0) + \epsilon_1$. Second, we coordinate the multi-resolution technique with the α refinement to optimally show the local information under specific α settings. The multi-resolution stage is terminated when further division no longer adds detail. All these operations are precomputed along with the projection without hampering real time performance, as the projection is the major time overhead for the visualization.

5 VISUALIZATION TOOLS

To support multi-platform development (DC6), we develop 1) a web-based visualization system, and 2) a computational notebook interface with a Python package to support visual exploration of multimodal embeddings.

5.1 Web-Based Visualization System

The web-based visualization system is implemented with a D3.js as the front end and a Python Flask as the back end. As shown in Fig. 6, the system consists of four views: (a) Settings Panel, (b) Overview, (c) Keyword Distribution View, and (d) Instance View.

Settings Panel. This panel (Fig. 6 (a)) allows users to select the dataset and cross-modal alignment metric of interest and displays relevant details about the chosen dataset and metric.

Overview. After selection, the view (Fig. 6 (b)) presents the corresponding map, which is the major component of our system featuring our *DKMap* method. This view consists of three modes: scatterplot mode, contour mode, and overlay mode.

- **Scatterplot mode.** The scatterplot mode displays all individual data points within the embedding space, with each point color-coded according to the cross-modal embedding metric. This mode allows for direct interaction with the data points.
- **Contour mode.** The contour mode can more effectively represent cross-modal metric distribution in a continuous manner. This mode makes regions with extreme values more prominent and reveals distribution patterns with greater clarity.
- **Overlay mode.** To enable users to simultaneously observe the overall distribution and interact with individual data points, we implement the overlay mode, which displays sampled data points on top of the contour map with different sampling methods of Blue noise sampling and random sampling, with dynamically increasing number of sample points as users zoom in.

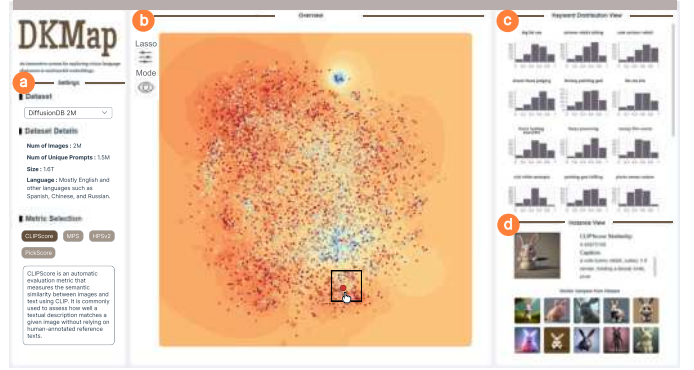


Fig. 6: **Web-based visualization system** consists of four views: (a) Settings Panel, (b) Overview presenting the projection results, (c) Keyword Distribution View, and (d) Instance View.

Keyword Distribution View. To enhance users’ interpretability of local subsets of data, we implement a keyword distribution view (Fig. 6 (c)) to show embeddings distribution in relation to interpretable textual features. Users can define a local subset via lasso selection on the map in the Overview, which dynamically constructs the corresponding embedding subset. We perform text concept keyword extraction on the subset. Specifically, we use KeyBERT [19] to extract keyword phrases from prompts and compute the cosine similarity between each keyword embedding and all text embeddings within the selected subset. These histograms reveal the distribution patterns of keywords within each group, highlighting variations in relevance and potential topic structures. Additionally, this view supports interactive exploration between the keyword distribution and the Overview. When the selected subset contains a large number of data points, users can narrow down their exploration by clicking bars in the keyword histogram to filter points in the Overview, enabling a focused analysis.

Instance View. The view (Fig. 6 (d)) complements the Overview by allowing users to interact with individual points in the embedding space. When users click the point, the panel below dynamically updates to display the corresponding image information. Additionally, the system employs Faiss [10] to retrieve and display similar images to the selected one, enhancing the user’s ability to explore related data rapidly.

5.2 Computational Notebook Interface

The Python package mainly supports the projection and contour mapping as well as the various multi-scale interactions of the Overview, which is implemented with dependencies on Pytorch and Plotly, and can be seamlessly integrated into interactive computational notebook.

Pytorch Projection. Here we illustrate some default settings of our Pytorch implementation. The architecture of our projection network is a 4-layer MLP, with shape (d, d) in the first three layers and $(d, 2)$ in the last layer, where d is the dimensions of the input embeddings. Batch normalization and ReLU activation are applied to each layer. These hyperparameters can be adjusted by users on demand.

Plotly Rendering. In the Plotly-based implementation of the visualization and interaction, we achieve the same three modes of Overview as in the web-based system. Particularly, we add dynamic zooming and sampled points overlay features, which are not available in original Plotly. Specifically, as Plotly is not optimized for interactive visualization, particularly for the large-scale embeddings in our scenario, we perform the following optimization techniques to support these interactions smoothly for large datasets and complex contour maps. For dynamic zooming, we implement a local update approach, where only the contour within the drawn region is re-rendered by the dynamically updated resolution and refined kernel parameter. For dynamic overlay, we implement a hierarchical tiling and progressive sampling approach. In this approach, we utilize the quad-tree node in the previous multi-resolution division as tiles to render local sample points. We hierarchically assign each point to a tile. Then, when the zoom region matches the size of tiles in a particular resolution, we find all the overlapping tiles and render all within the region sample points

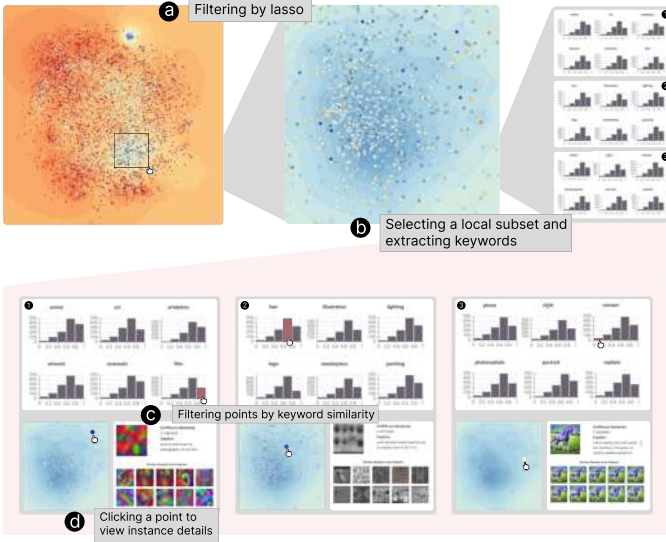


Fig. 7: **Exploring model-generated data quality with *DKMap*.** *DKMap* enables multi-faceted and multi-scale analysis with filtering by lasso in the Overview, filtering by features in the Keyword Distribution View, and selecting individual data points.

belonging to the tiles. This process is incremental, keeping all the previously shown points at higher level tiles.

6 EVALUATION

We conduct quantitative experiments (Sec. 6.1) and describe usage scenarios (Sec. 6.2) to demonstrate the effectiveness of *DKMap*.

6.1 Quantitative Experiments

Datasets. We conduct quantitative experiments on a commonly used dataset for alignment metric evaluation: the Pick-a-pic dataset [24].

Metrics. We compare *DKMap* against baseline methods for visualizing several commonly used cross-modal embedding metrics in T2I generation, including CLIPScore [21], HPSv2 [60], PickScore [24], and MPS [73]. In the quantitative experiments, we do not use any kernel cut-off technique as we aim to faithfully and fairly reflect the inherent properties of all kernels including the traditional Gaussian kernel and our dynamic kernel. As our method is a supervised DR method [54] that focuses on specific objectives (e.g., class separation in traditional supervised DR and mapping accuracy in our case), neighborhood preservation as in traditional unsupervised DR is not our main objective. We evaluate the accuracy of the visualization methods using mapping errors, including *mae*, *rmse*, and *mape*. These errors are computed at positions of projected points between the estimated value by the contour and the ground-truth alignment metric value of the particular point. Specifically, we calculate these errors on the test set for out-of-sample points that were not used during training. To show the trade-off between our main objective of mapping accuracy and traditional neighborhood preservation, we also provide statistics on neighborhood trustworthiness in the supplemental.

Baselines and Experiment Set-up. We compare *DKMap* against 1) three traditional DR methods: PCA with the Python sklearn package, t-SNE using sklearn with the Barnes-Hut t-SNE [50] implementation, and UMAP with the Python umap-learn package, 2) an encoder-based Neuro-Visualizer [11] implemented on the open-sourced code of the original paper [11], and 3) an anchor-based projection for multimodal embeddings - ModalChorus [71]. For the encoder-based methods, the decoder is used to estimate grid colors, while the traditional DR methods use two types of commonly adopted contour map estimation approaches: Gaussian RBF kernel (KDE) and Inverse Distance Weighting (IDW) [27] to render the 2D distribution map. As mentioned in the metrics part above, for all methods, the mapping accuracy is computed on the test set of Pick-a-pic, which contains 14,640 text-image pairs. When learning the projection, all the neural-network-based methods

(ours, Neuro-Visualizer and ModalChorus) are trained on Nvidia L4 device on Pick-a-pic’s training set which consists of 623,694 text-image pairs, while the traditional DR methods are computed on the union of train and test points as they are unsupervised and nonparametric. The t-SNE uses a perplexity parameter of 150 and the UMAP uses a neighborhood parameter of 50 to account for the large scale dataset. Both the ModalChorus and Neuro-Visualizer uses a training epoch of 1000 as the original papers of these methods require long epochs. In contrast, our method is trained for only 20 epochs by Adam optimizer with a learning rate of 0.002 and a batch size of 1000, and w_1 is set to 0.25.

Results and Analysis. The quantitative results in Tab. 1 reveal the advantages of our method consistently across all embeddings on the three accuracy metrics. Overall, *DKMap* significantly reduces the error by 30% - 90% compared to the best performing baseline. We also find that recent methods (e.g., Neuro-Visualizer and ModalChorus) do not work as well as traditional DR, with auto-encoder method performing the worst. This is because ModalChorus, as an anchor-based approach, cannot fuse large number of text and image embedding pairs in a unified point representation, while the auto-encoder-based approach fails to reconstruct complex multimodal embeddings. Both approaches are not suitable for revealing the distribution of alignment metric.

Figure 8 shows a qualitative comparison of some representative methods in both contour mode and scatterplot mode. Both t-SNE (KDE) and our method can produce smooth contour map, while ModalChorus and Neuro-Visualizer have obvious unsmoothness and inaccuracy. The poor contour map performance of Neuro-Visualizer comes from the difficulty for auto-encoder to reconstruct complex multimodal embeddings. The problem of ModalChorus is less serious, with large blocks on the periphery due to uneven density between central and border regions. The problem for traditional DR result is that the contour map lacks local detail. In comparison, under the same resolution, our *DKMap* can accurately estimate significantly more local detail. In addition, when we compare the scatterplot mode, we can also find that *DKMap* almost consistently outperforms other methods in clearly showing the metric distribution. For example, we can see the extreme values more prominently in *DKMap*. Although t-SNE seems to show the low extreme values quite well, it is not as good as *DKMap* for showing high extreme values in CLIPScore, HPSv2, and PickScore.

Additional experimental results are provided in the supplemental for more comprehensive evaluation, including in-sample mapping accuracy for training point positions and a user study. To show the efficiency of our package, in the supplemental we also record comparison with two commonly used python packages for dimensionality reduction: scikit-learn t-SNE and umap-learn in terms of running time on large-scale multimodal embeddings.

6.2 Usage Scenarios

This section presents three usage scenarios to demonstrate *DKMap*’s effectiveness in multi-scale exploration of vision-language alignment.

6.2.1 Scenario 1: Generated Dataset Exploration

Large-scale generated datasets such as DiffusionDB [56] provide important resources to help users understand T2I generated data and find references when crafting their own prompts [14]. However, previous visualization approaches to explore these datasets suffer from modality gap [55] or scalability problem [14], which cannot visualize the generative performance across the whole dataset. Our *DKMap* can address this problem by visualizing the whole million-scale large dataset together with the alignment score distribution.

As illustrated in Fig. 7, users first begin by selecting a region of interest using a lasso filter in the Overview (Fig. 7(a)). Here, the user is interested in an area where alignment scores are relatively low. After selection, the system will automatically extract keywords (Fig. 7(b)) from the prompts within this subset. Each extracted keyword is associated with a histogram, where the bars represent different similarity ranges between the keyword’s embedding and the text embeddings of the prompts in the subset. With the feature distribution, users can further adjust their selection by filtering data points based on keyword

Table 1: Quantitative comparison of cross-modal embedding metric contour mapping accuracy on the Pick-a-pic dataset [24]. For all metrics, the lower the better. Our method outperforms baseline methods, achieving the best performance across all four cross-modal embedding metrics.

Method	CLIPScore			HPSv2			MPS			PickScore		
	mae ↓	mape ↓	rmse ↓	mae ↓	mape ↓	rmse ↓	mae ↓	mape ↓	rmse ↓	mae ↓	mape ↓	rmse ↓
PCA (KDE)	2.898	9.966	3.777	1.505	5.476	1.903	3.105	679.623	3.880	1.099	5.322	1.384
PCA (IDW)	3.052	10.502	3.956	1.556	5.669	1.973	3.172	776.244	4.013	1.141	5.527	1.443
t-SNE (KDE)	2.935	10.113	3.818	1.370	5.006	1.757	3.025	512.417	3.694	1.033	4.997	1.290
t-SNE (IDW)	2.728	9.327	3.569	1.323	4.829	1.703	2.754	796.972	3.497	0.951	4.585	1.207
UMAP (KDE)	2.931	10.053	3.793	1.599	5.838	2.027	3.562	1150.590	4.596	1.138	5.495	1.430
UMAP (IDW)	2.663	9.149	3.524	1.412	5.167	1.840	3.174	1101.555	4.176	1.065	5.120	1.354
Neuro-Visualizer	9.659	30.137	11.008	3.385	11.810	4.038	4.608	515.580	5.494	2.197	10.218	2.680
ModalChorus	3.270	11.304	4.187	1.634	5.948	2.065	3.616	1150.937	4.703	1.101	5.288	1.383
DKMap (w/o KR)	3.116	10.501	4.004	1.451	5.262	1.855	3.025	852.597	3.765	1.002	4.832	1.263
DKMap	1.692	5.570	2.231	0.895	3.224	1.151	0.861	71.465	1.122	0.717	3.493	0.936

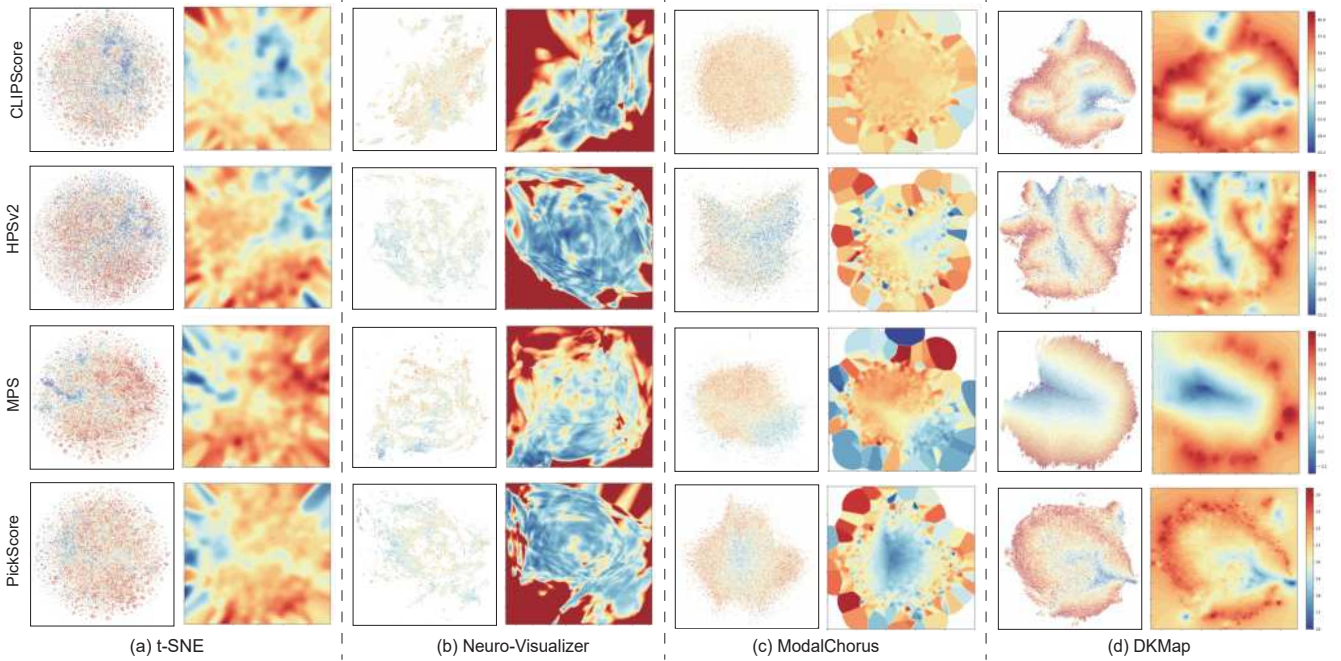


Fig. 8: **Qualitative comparison of DKMap vs. baseline methods:** (a) t-SNE [51], a traditional DR, (b) Neuro-Visualizer [51], an auto-encoder method, (c) ModalChorus [71], a recent anchor-based method, and (d) our DKMap. DKMap presents more evident metric distribution in scatterplot mode and more accurate contour map in contour mode.

similarity. For instance, the user selects the keyword “film” and chooses data points with a similarity score between 0.8 and 1.0 (Fig. 7(c1)). Simultaneously, the Overview updates by highlighting only the prompts strongly related to “film” while filtering out the rest, resulting in instances most related to “film” (Fig. 7(d1)). Similarly, the user can filter features of the other keyword “hair” (Fig. 7(c2)) and examine related data points (Fig. 7(d2)). By iteratively selecting different keywords and similarity ranges, users can identify patterns in low alignment score prompts and find better references for crafting improved prompts.

6.2.2 Scenario 2: Generative Model Comparison

DKMap facilitates comparative evaluation of generative models through two distinct modes: 1) side-by-side comparison that visualizes multimodal embeddings from different models in adjacent views; and 2) explicit encoding by explicitly concatenating the multimodal embeddings, followed by visualizing the concatenated embeddings in one unified view. While side-by-side comparison is common in existing methods (e.g., [55]), DKMap uniquely supports explicit encoding for more rigorous generative model comparison. Here, we compare two different foundation models: SDXL [42] and SD3-medium [13] on the COCO validation set consisting of over 25000 prompts.

As shown in Fig. 9, we use contour map to visualize the global distribution of the HPSv2 alignment score for the two models. Figure 9(a) first shows the embedding distributions for individual models

side-by-side. While this side-by-side visualization reveals each model’s overall structure, the inherent instability of DR process prevents direct point-to-point comparison between models, limiting detailed analysis of individual correspondences. Alternatively, by concatenating the image embeddings of the two models’ results with the prompt text embeddings and projecting the data using DKMap, we can visualize the distribution of the explicit difference between these models as in Fig. 9(b). Through this visualization, some interesting findings are revealed. First, even though Stable Diffusion 3 is a more recent model than SDXL with the new Diffusion Transformer architecture, it does not always outperform the SDXL model, as both positive and negative differences between their alignment performance exist. For example, SDXL tends to generate indoor scenes with higher HPSv2 scores as in Fig. 9(b-2), where SDXL typically generates a more complete scene. In comparison, SD3 beats SDXL in other regions, such as some wild animal generation in Fig. 9(b-1), where SD3 typically generates clearer and brighter images. As such, explicit encoding enables a more direct and intuitive comparison.

6.2.3 Scenario 3: Billion-Scale Pretraining Visualization

In the age of large models, tremendous-scale pretraining is prevalent. For example, multimodal pretraining datasets such as LAION [46] and DataComp [15] already reach the size close to one billion. Visualizing multimodal embeddings at such a scale presents a significant challenge.

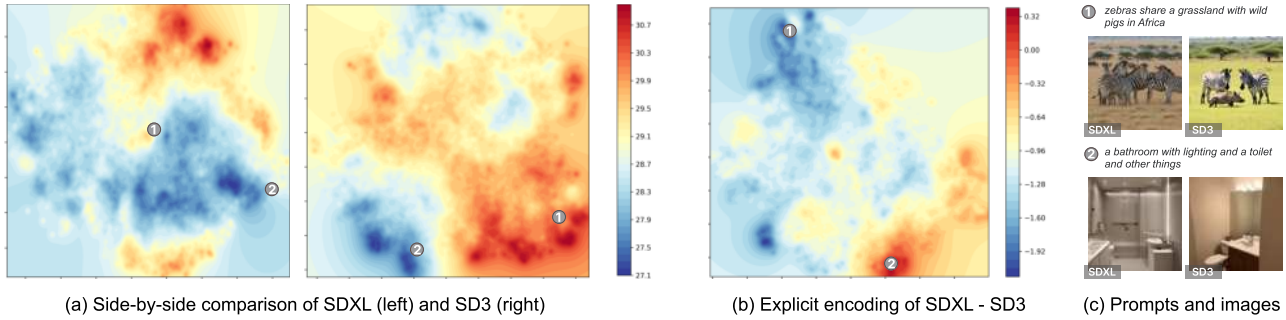


Fig. 9: **Comparing generative models with *DKMap*.** *DKMap* enables (a) side-by-side comparison and also (b) explicit encoding of vision-language alignment in different models. (c) Instance-level comparison shows text-image pairs generated by both models for the same input prompt.

In this case, we show that *DKMap* has excellent scalability in handling such huge datasets, with an example of visualizing CLIP embeddings of LAION-400M [46], which is 200 times larger than diffusionDB-2M [56] as in case 1 and more than 1000 times larger than COCO [32].

Figure 1 shows the visualization of LAION-400M in the output cell of Colab notebook. In this case, we use the overlay mode of our visualization tool so that users can simultaneously see the distribution of CLIPScore and the sample points. Particularly, the multi-scale exploration makes it possible to display all the data points through hierarchical dynamic overlay, which is otherwise impossible to render in a static overview for 400 million points. As shown in Fig. 1(left), it is difficult to interact with large number of points in the highest level overview, where the contour map plays a more important role in revealing aggregate information about metric distribution. By dynamic zooming, users can focus on specific region of interest as in the middle, where our method dynamically reveals more fine-grained distribution in the contour map and increases sampled points to show details and outliers. Users can further perform multiple rounds of zooming to incrementally show even more enhanced detail in a hierarchical manner. Finally, users can mouse over particular points of interest to show instance-level data including image, caption, url and CLIPScore. Such a point would be otherwise highly difficult to render, find and interact with in a static scatterplot of such scale. Note that *DKMap* enables visualization of datasets of such scale in computational notebook without relying on special GPU-accelerated rendering library (e.g., WebGL).

7 DISCUSSION

Application to unimodal embedding. Although in this work we focus on multimodal embedding mapping, our dynamic kernel approach can be feasibly extended as a general framework for accurate and multi-scale mapping of any high-dimensional embedding-derived metrics. The strategy of jointly learning discrete point DR with continuous metric landscape estimation proves highly effective in many scenarios. For example, we demonstrate in the supplementary material a scenario of visualizing unimodal image Aesthetic Score, where our method can also improve upon traditional DR + kernel mapping method.

Application to multimodal embedding beyond vision-language. Multimodal embeddings have evolved beyond vision-language alignment to incorporate diverse modalities. Recent advances such as ImageBind [16] and LanguageBind [77] demonstrate successful integration of additional modalities (e.g., audio, video, thermal, and depth) with visual and language representations. While our framework demonstrates promising generalizability for embedding-based metric visualization across tasks such as 3D and video generation, the current implementation may face challenges when processing complex spatiotemporal modalities such as video. Nevertheless, as vision-language alignment is the basis of multimodal learning, our method has strong potential to generalize to other modalities. For example, other cross-modal tasks such as text-to-3D [30], text-to-video [34] and text-to-SVG [58] also rely on similar metrics such as CLIPScore in the evaluation, which provide broader potential applications of our visualization.

Embedding-based human-AI interaction. The rise of large vision-language and generative models introduces new opportunities and challenges for human-AI interaction. Unlike earlier models focused on

narrow tasks, these foundation models can handle complex, high-level problems, prompting users to delegate difficult tasks to them. In visualization and HCI, this shift has led to growing adoption and a push for intuitive interfaces that harness model capabilities while abstracting backend complexity [63, 64, 67]. Despite leveraging the rich knowledge in pretrained models, current systems offer limited human control beyond prompting. Embedding representations provide a promising path for deeper human intervention, as they encode the model’s learned knowledge. Recent training-free methods, such as embedding manipulation via cross-attention [23], allow user-guided generation, but a gap remains between neural embeddings and human-understandable concepts, hindering precise control. To enable effective embedding-based interaction, three challenges must be addressed: 1) exploring large embedding spaces, 2) interpreting embedding semantics, and 3) fine-tuning model behavior. This work focuses on the first challenge, developing methods for scalable embedding exploration and analysis.

Limitations and future work. While *DKMap* effectively evaluates direct vision-language alignment in pretrained embeddings, it faces challenges in indirect evaluation of pretrained embeddings’ quality. Current evaluation paradigms rely primarily on downstream metrics such as zero-shot retrieval recall [25] rather than direct analysis of cross-modal distances derived from embeddings. This limitation stems from our method’s current inability to fully characterize the structural properties of multimodal embedding spaces, particularly their relationship to underlying dataset characteristics. We envision addressing this gap by developing principled approaches for embedding-space analysis that complement distance-based metrics. Moreover, to enable more interpretable human-AI interaction, we plan to develop enhanced methods for analyzing multimodal representations. We will focus on uncovering implicit knowledge encoded in the embeddings, translating these representations into human-interpretable concepts, and characterizing how these concepts influence model behavior [36, 41].

8 CONCLUSION

In this study, we present *DKMap*, a multi-scale interactive visualization technique for exploring vision-language alignment in multimodal embeddings. *DKMap* enables accurate contour mapping of alignment metrics at scale and supports dynamic zooming through kernel-enhanced projection. It bridges dimensionality reduction and post-projection mapping via kernel regression-supervised t-SNE, and enhances local detail with a multi-resolution kernel refinement technique. We implement *DKMap* as both a web-based system and a computational notebook interface for accessible and flexible exploration. Quantitative evaluations show improved mapping accuracy over common DR, autoencoder-based, and anchor-based methods. We demonstrate its utility through several use cases: exploring million-scale DiffusionDB, comparing T2I models, and interacting with billion-scale LAION data. Code and demos are available at <https://github.com/HKUST-CIVAL/DKMap>.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (62172398, 62402184), Guangxi Science and Technology Program (25069470), and the Guangdong Provincial Fund for

REFERENCES

- [1] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010. doi: 10.1002/wics.101 1, 2
- [2] A. Boggust, B. Carter, and A. Satyanarayan. Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples. In *Proc. ACM IUI*, pp. 746–766. ACM, New York, 2022. doi: 10.1145/3490099.3511122 3, 4
- [3] C. Chen, F. Lv, Y. Guan, P. Wang, S. Yu, Y. Zhang, and Z. Tang. Human-guided image generation for expanding small-scale training image datasets. *IEEE Trans. Vis. Comput. Graph.*, 31(6):3809–3821, 2025. doi: 10.1109/TVCG.2025.3567053 2, 3
- [4] J. Chen, Q. Huang, C. Wang, and C. Li. SenseMap: Urban performance visualization and analytics via semantic textual similarity. *IEEE Trans. Vis. Comput. Graph.*, 30(9):6275–6290, 2024. doi: 10.1109/TVCG.2023.3333356 3
- [5] 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, 2016. doi: 10.1109/TVCG.2015.2467552 3
- [6] Y. Cheng, X. Wang, and Y. Xia. Supervised t-distributed stochastic neighbor embedding for data visualization and classification. *INFORMS Journal on Computing*, 33(2):566–585, 2021. doi: 10.1287/ijoc.2020.0961 2, 3
- [7] C. Collins and S. Carpendale. VisLink: Revealing relationships amongst visualizations. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1192–1199, 2007. doi: 10.1109/TVCG.2007.70521 4
- [8] C. De Bodt, D. Mulders, M. Verleysen, and J. A. Lee. Perplexity-free t-SNE and twice student tt-SNE. In *Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 123–128. i6doc.com publ, Bruges, 2018. doi: 2078.1/200844 5
- [9] S. Demir and Ö. Toktamiş. On the adaptive nadaraya-watson kernel regression estimators. *Hacettepe Journal of Mathematics and Statistics*, 39(3):429–437, 2010. doi: 20.500.12809/4526 3
- [10] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024. doi: 10.48550/arXiv.2401.08281 6
- [11] M. Elhamod and A. Karpatne. Neuro-Visualizer: A novel auto-encoder-based loss landscape visualization method with an application in knowledge-guided machine learning. In *Proc. ICML*, pp. 12429–12447. PMLR, Vienna, 2024. 3, 7
- [12] M. Espadoto, G. Appleby, A. Suh, D. Cashman, M. Li, C. Scheidegger, E. W. Anderson, R. Chang, and A. C. Telea. Unprojection: Leveraging inverse-projections for visual analytics of high-dimensional data. *IEEE Trans. Vis. Comput. Graph.*, 29(2):1559–1572, 2021. doi: 10.1109/TVCG.2021.3125576 3
- [13] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. ICML*, pp. 12606–12633. PMLR, Vienna, 2024. 8
- [14] 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, 2023. doi: 10.1109/TVCG.2023.3327168 7
- [15] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *Proc. NIPS*, pp. 27092–27112. Curran Associates, Inc., New York, 2023. 8
- [16] 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. IEEE, New York, 2023. doi: 10.1109/CVPR52729.2023.01457 9
- [17] A. Gisbrecht, A. Schulz, and B. Hammer. Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing*, 147:71–82, 2015. doi: 10.1016/j.neucom.2013.11.045 3
- [18] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011. doi: 10.1177/1473871611416549 4
- [19] M. Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. doi: 10.5281/zenodo.4461265 6
- [20] 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 3, 4
- [21] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. EMNLP*, pp. 7514–7528. ACL, Punta Cana, 2021. doi: 10.18653/v1/2021.emnlp-main.595 1, 2, 4, 7
- [22] 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, Virtual, 2021. 3
- [23] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu. Dense text-to-image generation with attention modulation. In *Proc. ICCV*, pp. 7701–7711. IEEE, New York, 2023. doi: 10.1109/ICCV51070.2023.00708 9
- [24] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Proc. NIPS*, pp. 36652–36663. Curran Associates, Inc., New York, 2023. 2, 7, 8
- [25] Z. Lai, H. Zhang, B. Zhang, W. Wu, H. Bai, A. Timofeev, X. Du, Z. Gan, J. Shan, C.-N. Chuah, et al. VeCLIP: Improving clip training via visual-enriched captions. In *Proc. ECCV*, pp. 111–127. Springer, Milan, 2024. doi: 10.1007/978-3-031-72946-1_7 9
- [26] O. D. Lampe and H. Hauser. Interactive visualization of streaming data with kernel density estimation. In *Proc. IEEE PacificVis*, pp. 171–178. IEEE, Hong Kong, 2011. doi: 10.1109/PACIFICVIS.2011.5742387 4, 5
- [27] R. le Roux, S. Henrico, J. Bezuidenhout, and I. Henrico. Inverse distance weighting as an alternative interpolation method to create radiometric maps of natural radionuclide concentrations using qgis. In *Proceedings of the ICA*, vol. 5, p. 10. Copernicus Publications Göttingen, Germany, 2023. doi: 10.5194/ica-proc-5-10-2023 7
- [28] J. Li and J. Kuang. Robustmap: Visual exploration of dnn adversarial robustness in generative latent space. *IEEE Trans. Vis. Comput. Graph.*, pp. 1–15, 2024. doi: 10.1109/TVCG.2024.3471551 3
- [29] Q. Li, K. S. Njotoprawiro, H. Haleem, Q. Chen, C. Yi, and X. Ma. Embeddingvis: A visual analytics approach to comparative network embedding inspection. In *Proc. IEEE VAST*, pp. 48–59. IEEE, Berlin, 2018. doi: 10.1109/VAST.2018.8802454 3
- [30] X.-L. Li, H. Li, H.-X. Chen, T.-J. Mu, and S.-M. Hu. DiScene: Object decoupling and interaction modeling for complex scene generation. In *Proc. SIGGRAPH Asia*, pp. 1–12. ACM, Tokyo, 2024. doi: 10.1145/3680528.3687589 1, 2, 9
- [31] 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. In *Proc. NIPS*, pp. 17612–17625. Curran Associates, Inc., New Orleans, 2022. 2
- [32] 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, Zurich, 2014. doi: 10.1007/978-3-319-10602-1_48 9
- [33] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Proc. NIPS*, pp. 34892–34916. Curran Associates, Inc., New York, 2023. 1
- [34] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proc. CVPR*, pp. 22139–22149. Seattle, 2024. doi: 10.1109/CVPR52733.2024.02090 9
- [35] Y. Liu, E. Jun, Q. Li, and J. Heer. Latent space cartography: Visual analysis of vector space embeddings. *Comput. Graph. Forum*, 38:67–78, 2019. doi: 10.1111/cgf.13672 3
- [36] J. Materzyńska, A. Torralba, and D. Bau. Disentangling visual and written concepts in CLIP. In *Proc. CVPR*, pp. 16410–16419. IEEE, New Orleans, 2022. doi: 10.1109/CVPR52688.2022.01592 9
- [37] A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Trans. Vis. Comput. Graph.*, 19(9):1526–1538, 2013. doi: 10.1109/TVCG.2013.65 3
- [38] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426 1, 2
- [39] A. Narayan, B. Berger, and H. Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, 39(6):765–774, 2021. doi: 10.1038/s41587-020-00801-7 3, 5
- [40] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Trans. Vis. Comput. Graph.*, 25(8):2650–2673, 2018. doi: 10.

- [41] J. Parekh, P. Khayatan, M. Shukor, A. Newson, and M. Cord. A concept-based explainability framework for large multimodal models. In *Proc. NIPS*, pp. 135783–135818. Curran Associates, Inc., Red Hook, 2024. 9
- [42] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proc. ICLR*, pp. 1–13. Openreview.net, Vienna, 2024. 8
- [43] 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. PMLR, Virtual, 2021. 1, 2, 3
- [44] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. doi: 10.48550/arXiv.2204.06125 2
- [45] 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. New Orleans, 2022. doi: 10.1109/CVPR52688.2022.01042 1, 2
- [46] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proc. NIPS*, pp. 25278–25294. Curran Associates, Inc., New Orleans, 2022. 1, 2, 4, 8, 9
- [47] A. Schulz and B. Hammer. Discriminative dimensionality reduction for regression problems using the fisher metric. In *Proc. IJCNN*, pp. 1–8. IEEE, Killarney, 2015. doi: 10.1109/IJCNN.2015.7280736 2, 3
- [48] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018. doi: 10.1201/9781315140919 3, 5
- [49] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien. Linear discriminant analysis: A detailed tutorial. *AI Communications*, 30(2):169–190, 2017. doi: 10.3233/AIC-170729 4
- [50] L. Van Der Maaten. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014. doi: doi/10.5555/2627435.2697068 7
- [51] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008. 1, 2, 8
- [52] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Trans. Vis. Comput. Graph.*, 28(1):802–812, 2021. doi: 10.1109/TVCG.2021.3114794 3
- [53] X. Wang, R. Huang, Z. Jin, T. Fang, and H. Qu. Commonsensevis: Visualizing and understanding commonsense reasoning capabilities of natural language models. *IEEE Trans. Vis. Comput. Graph.*, 30(1):273–283, 2023. doi: 10.1109/TVCG.2023.3327153 3
- [54] Y. Wang, K. Feng, X. Chu, J. Zhang, C.-W. Fu, M. Sedlmair, X. Yu, and B. Chen. A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE Trans. Vis. Comput. Graph.*, 24(5):1828–1840, 2017. doi: 10.1109/TVCG.2017.2701829 2, 3, 7
- [55] Z. J. Wang, F. Hohman, and D. H. Chau. WizMap: Scalable interactive visualization for exploring large machine learning embeddings. In *Proc. ACL*, pp. 516–523. ACL, Toronto, 2023. doi: 10.18653/v1/2023.acl-demo.50 2, 3, 7, 8
- [56] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proc. ACL*, pp. 893–911. Toronto, 2023. doi: 10.18653/v1/2023.acl-long.51 4, 7, 9
- [57] S. Węglarczyk. Kernel density estimation and its application. In *ITM Web of Conferences: XLVIII Seminar of Applied Mathematics*, vol. 23, article no. 00037. EDP Sciences, 2018. doi: 10.1051/itmconf/20182300037 3, 5
- [58] J. Wei, C. Tan, Q. Chen, G. Wu, S. Li, Z. Gao, L. Sun, B. Yu, and R. Guo. From words to structured visuals: A benchmark and framework for text-to-diagram generation and editing. In *Proc. CVPR*, pp. 13315–13325. IEEE, New York, 2025. 9
- [59] D. M. Witten and R. Tibshirani. Supervised multidimensional scaling for visualization, classification, and bipartite ranking. *Computational Statistics & Data Analysis*, 55(1):789–801, 2011. doi: 10.1016/j.csda.2010.07.001 2
- [60] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. doi: 10.48550/arXiv.2306.09341 1, 2, 4, 7
- [61] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li. Human preference score: Better aligning text-to-image models with human preference. In *Proc. ICCV*, pp. 2096–2105. IEEE, New York, 2023. doi: 10.1109/ICCV51070.2023.00200 2
- [62] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis. Linear discriminant analysis. *Robust Data Mining*, pp. 27–33, 2013. doi: 10.1007/978-1-4419-9878-1 3
- [63] S. Xiao, S. Huang, Y. Lin, Y. Ye, and W. Zeng. Let the chart spark: Embedding semantic context into chart with generative model. *IEEE Trans. Vis. Comput. Graph.*, 30(1):284 – 294, 2024. doi: 10.1109/TVCG.2023.3326913 9
- [64] S. Xiao, L. Wang, X. Ma, and W. Zeng. TypeDance: Creating semantic typographic logos from image through personalized generation. In *Proc. ACM CHI*, article no. 175, pp. 1–18. ACM, New York, 2024. doi: 10.1145/3613904.364218 9
- [65] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Proc. NIPS*, pp. 15903–15935. Curran Associates, Inc., Red Hook, 2023. 2
- [66] W. Yang, M. Liu, Z. Wang, and S. Liu. Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media*, 10(3):399–424, 2024. doi: 10.1007/s41095-023-0393-x 2
- [67] 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. *Vis. Inf.*, 8(2):43–66, 2024. doi: 10.1016/j.visinf.2024.04.003 9
- [68] Y. Ye, J. Huang, J. Xia, W. Zeng, et al. AKRMap: Adaptive kernel regression for trustworthy visualization of cross-modal embeddings. In *Proc. ICML*. PMLR, Virtual, 2025. 4
- [69] 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.*, 30(7):3224–3240, 2022. doi: 10.1109/TVCG.2022.3229023 3
- [70] Y. Ye, R. Huang, K. Zhang, and W. Zeng. Unified visual comparison framework for human and ai paintings using neural embeddings and computational aesthetics. *IEEE Comput. Graph. Appl.*, 45(2):19–30, 2025. doi: 10.1109/MCG.2025.3555122 3
- [71] Y. Ye, S. Xiao, X. Zeng, and W. Zeng. Modalchorus: Visual probing and alignment of multi-modal embeddings via modal fusion map. *IEEE Trans. Vis. Comput. Graph.*, 31(1):294–304, 2024. doi: 10.1109/TVCG.2024.3456387 2, 3, 7, 8
- [72] 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. *Proceedings of the ACM on Human-Computer Interaction*, 8(180):1–31, 2024. doi: 10.1145/364101 1, 2
- [73] S. Zhang, B. Wang, J. Wu, Y. Li, T. Gao, D. Zhang, and Z. Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proc. CVPR*, pp. 8018–8027. IEEE, New York, 2024. doi: 10.1109/CVPR52733.2024.00766 7
- [74] T. Zhang, J. Li, and C. Xu. Visual exploration of multi-dimensional data via rule-based sample embedding. *Vis. Inf.*, 8(3):53–56, 2024. doi: 10.1016/j.visinf.2024.09.005 3
- [75] 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 3
- [76] 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 *Proc. SIGGRAPH*, article no. 57, pp. 1–9. ACM, New York, 2023. doi: 10.1145/3588432.359153 2, 3
- [77] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. doi: 10.48550/arXiv.2310.01852 9