# Semantic-Space-Intervened Diffusive Alignment for Visual Classification

**Zixuan Li**[1] , **Lei Meng**[1*] , **Guoqing Chao**[2] , **Wei Wu**[1] , **Yimeng Yang**[1] ,
**Xiaoshuo Yan**[1] , **Zhuang Qi**[1] , **Xiangxu Meng**[1]
[1]School of Software, Shandong University, Jinan, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China
{lizixuan0707, wu_wei, yanxiaoshuo, y_yimeng, z_qi }@mail.sdu.edu.cn,
{lmeng, mxx}@sdu.edu.cn, guoqingchao@hit.edu.cn

## Abstract

Cross-modal alignment is an effective approach to improving visual classification. Existing studies typically enforce a one-step mapping that uses deep neural networks to project the visual features to mimic the distribution of textual features. However, they typically face difficulties in finding such a projection due to the two modalities in both the distribution of class-wise samples and the range of their feature values. To address this issue, this paper proposes a novel **Se**mantic-Space-Intervened **D**iffusive **A**lignment method, termed *SeDA*, models a semantic space as a bridge in the visual-to-textual projection, considering both types of features share the same class-level information in classification. More importantly, a bi-stage diffusion framework is developed to enable the progressive alignment between the two modalities. Specifically, SeDA first employs a Diffusion-Controlled Semantic Learner to model the semantic features space of visual features by constraining the interactive features of the diffusion model and the category centers of visual features. In the later stage of SeDA, the Diffusion-Controlled Semantic Translator focuses on learning the distribution of textual features from the semantic space. Meanwhile, the Progressive Feature Interaction Network introduces stepwise feature interactions at each alignment step, progressively integrating textual information into mapped features. Experimental results show that SeDA achieves stronger cross-modal feature alignment, leading to superior performance over existing methods across multiple scenarios.

## 1 Introduction

Cross-modal alignment aims to integrate information from different modalities to capture semantic relationships within complex data [Baltrušaitis *et al.*, 2019; Dang *et al.*, 2024b]. It utilizes more discriminative textual representations to enhance visual classification, effectively mitigating biases caused by the diversity of visual data, lighting conditions, and background noise [Dang *et al.*, 2025; Dang *et al.*, 2023; Dang *et al.*, 2024c]. Although cross-modal alignment generally outperforms single-modal learning, its effectiveness can decline due to semantic ambiguities between modalities [Meng *et al.*, 2019; Dang *et al.*, 2024a]. This is primarily due to substantial differences in semantics, structures, or representational forms across modalities, which pose significant challenges for cross-modal alignment in handling high heterogeneity.

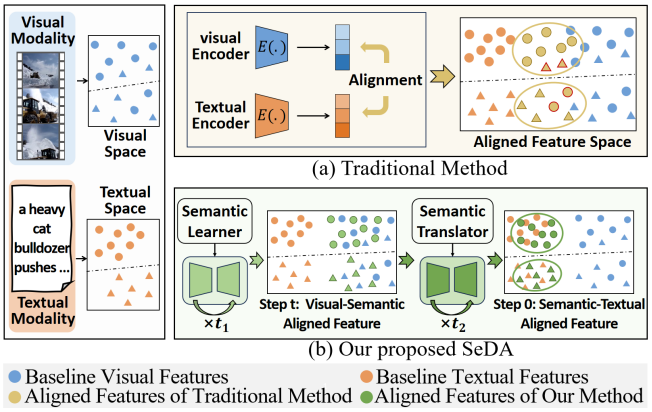


Figure 1: Common feature alignment methods and the proposed SeDA alignment framework. In (a), traditional feature alignment processes fail to capture the underlying distribution of textual features, resulting in persistent inter-class confusion. In (b), SeDA employs a semantic-space-intervened diffusive alignment method, transferring visual features to the textual features space step by step through a bi-stage learning process.

To alleviate modality heterogeneity, existing methods can be broadly divided into two groups: distance metric-based alignment methods and knowledge distillation-based alignment methods. The former approach uses techniques like Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2012], Correlation Alignment (CORAL) [Sun *et al.*, 2016], and the Wasserstein Distance [Lee *et al.*, 2019; Deng *et al.*, 2025] to explicitly minimize the differences in feature distributions or representations between modalities [Fu *et al.*, 2025]. The latter approach leverages [Huo *et al.*, 2024; Aslam *et al.*, 2023; Xu *et al.*, 2024] to transfer knowledge from the teacher modality to the student modality, enabling different modalities to generate similar feature representations. However, as shown in Figure 1(a), due to the significant differences in

*Corresponding author

class-wise sample distributions and feature value ranges between modalities, directly applying a one-step mapping often fails to find an effective projection, failing to capture the underlying distribution of textual features fully. Furthermore, the complex relationships between modalities make it challenging for one-step alignment methods to adequately address modality-specific distributions that are highly correlated with each modality's intrinsic discriminability. Consequently, semantic ambiguity remains, leading to significant inter-class confusion in the aligned features.

To address the above challenges, we propose a **Se**mantic-Space-Intervened **D**iffusive **A**lignment (SeDA) method, as illustrated in Figure 1(b). SeDA leverages the Markov reverse process of diffusion models to smoothly learn distributions over multiple steps, focusing on the semantic consistency of multimodal features. This approach alleviates the semantic ambiguity caused by the heterogeneity of textual and visual features, which arises from direct one-step mapping. SeDA consists of three key modules: the Diffusion-controlled Semantic Learner (DSL), the Diffusion-controlled Semantic Translator (DST), and the Progressive Feature Interaction Network (PFIN). Specifically, we propose a Bi-stage optimization framework that models a modality-shared semantic space as an intermediary to enable a three-stage projection: from visual space to semantic space and then to textual space. In the early stage of the diffusion process, the DSL module progressively removes irrelevant low-level visual information by regularizing the distance between the features learned from the PFIN module and the category center of the original visual features, thereby modeling semantic space from visual representations. In the later stage, the DST module guides the transformation of semantic space to the textual representation space by measuring the distributional differences between semantic and textual features. Additionally, the PFIN module designs a diffusion network based on a cross-attention mechanism for multimodal feature fusion, ensuring the gradual introduction of textual information into mapped features through interaction at each diffusion step. SeDA not only ensures similarity between the mapped feature distribution and the textual distribution but also considers modality-independent semantic information in both visual and textual representations.

Extensive experiments are conducted on the general dataset NUS-WIDE, the domain-specific dataset VIREO Food-172, and the video dataset MSRVTT, including performance comparisons, ablation studies, in-depth analysis, and case studies. The experimental results show that SeDA models a semantic space as a bridge for the visual-to-textual projection, alleviating modality heterogeneity and achieving the transformation from visual features to textual features. The contributions of this paper are as follows:

- A novel framework SeDA is proposed to enable the alignment between visual and textual features by modeling a diffusion process. To the best of our knowledge, SeDA is the first work to use diffusion models for cross-modal alignment in classification.
- The developed diffusion process improves the visual-to-textual feature projection by modeling a semantic space, which may capture higher-level semantic relationships between visual and textual representations, serving as an "intermediary layer" that effectively reduces the heterogeneity between different modalities.
- SeDA is a model-agnostic framework that can integrate into various visual backbones. It effectively learns the underlying distribution of textual features, providing a feasible approach for future research.

## 2 Related Work

### 2.1 Cross-Modal Alignment

**Distance Metrics-based Alignment Methods** focus on minimizing or maximizing metrics to bring data from different modalities into a common feature or decision space. Common metrics include Euclidean distance, cosine similarity, and covariance differences. For instance, Coral [Sun *et al.*, 2016] aligns source and target features by minimizing the Frobenius norm of their feature differences. Deep Coral [Sun and Saenko, 2016] extends this by incorporating decision-level information and utilizing covariance matrices as a new alignment metric. Similarly, CLIP [Radford *et al.*, 2021], ECRL [Wang *et al.*, 2024], and TEAM [Xie *et al.*, 2022] enhance cross-modal alignment by computing cosine similarity between visual and textual representations, effectively bridging the gap between modalities in the shared space.

**Knowledge Distillation-based Alignment Methods** transfer knowledge from one modality to another by employing a pre-trained teacher model to guide a student model. This ensures that both modalities generate similar representations in a shared alignment space. For example, C2KD [Huo *et al.*, 2024] utilizes bidirectional distillation and dynamically filters samples with misaligned soft labels to improve alignment. MM-PKD [Aslam *et al.*, 2023] employs a multimodal teacher network to guide an unimodal student network through joint cross-attention fusion. Furthermore, PKDOT [Aslam *et al.*, 2024] leverages entropy-regularized optimal transport to distill structural knowledge, enhancing stability and robustness in the multimodal distillation process.

### 2.2 Diffusion Models

Diffusion models are inspired by non-equilibrium thermodynamics. DDPM [Ho *et al.*, 2020] gradually adds noise to the data distribution and trains a neural network to learn the reverse diffusion process, thereby denoising images corrupted by Gaussian noise. Most diffusion model studies focus on generative tasks, such as image generation [Wang *et al.*, 2023], 3D content generation [Li *et al.*, 2024], and video generation [Ho *et al.*, 2022]. Recently, some work has applied diffusion models to discriminative tasks [Yang *et al.*, 2025]. For example, DiffusionDet [Chen *et al.*, 2023] formulates object detection as a diffusion denoising process, progressively refining noisy bounding boxes into object boxes. Diffusion-Ret [Jin *et al.*, 2023] models the correlation between text and video as their joint probability and approaches retrieval as a gradual generation of this joint distribution from noise. DiffuMask [Wu *et al.*, 2023] uses text-guided cross-attention information to locate class- or word-specific regions, resulting in semantic masks for synthesized images.

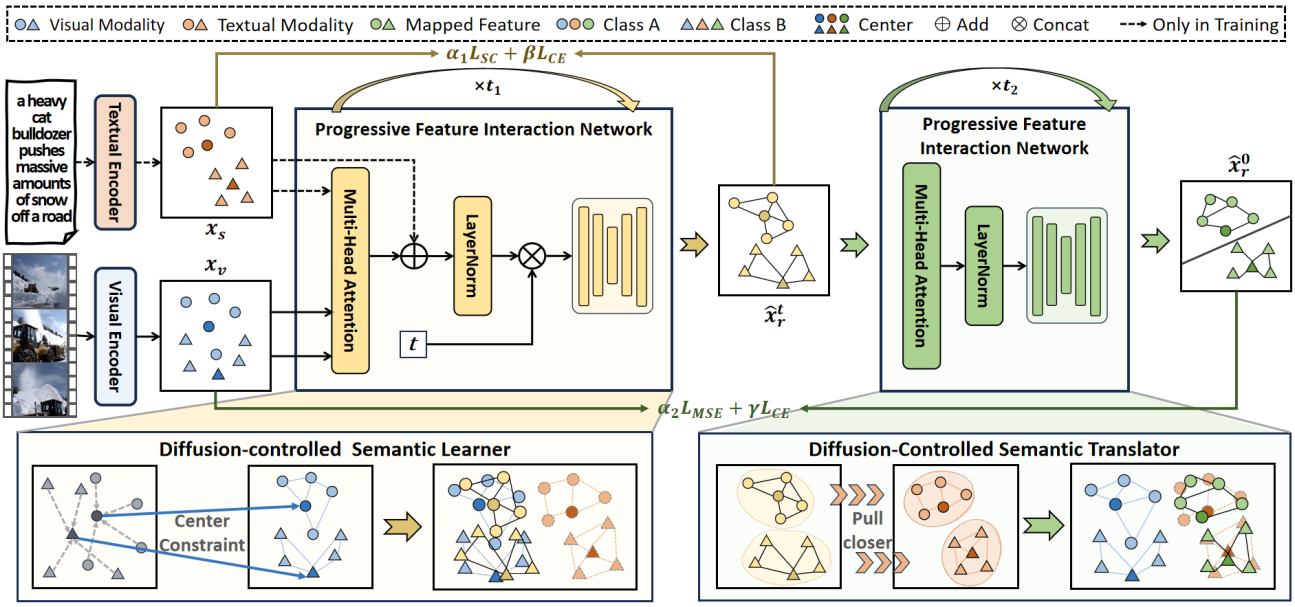


Figure 2: Illustration of the proposed SeDA. SeDA takes visual-textual data pairs as input, which are processed by dedicated neural networks for vision and text to extract global features $x_v$ and $x_s$. The PFIN module progressively integrates textual information, while the DSL and DST modules work together to align visual and textual features effectively.

## 3 Problem Formulation

The goal of this research is to improve the classification performance of cross-modal alignment networks by leveraging learning using privileged information (LUPI), where textual data is available only during the training phase. The dataset is composed of paired visual data $V$ and textual data $T$, with the rare but informative textual data $T$ serving as privileged information to enhance the representation of visual data $V$. Specifically, the training set consists of $N$ visual-textual pairs $D_N = (v_1, t_1), (v_2, t_2), ..., (v_n, t_n)$, while the test set contains only $M$ visual samples $D_M = v_1, v_2, ..., v_m$. A visual encoder $E_v$ is employed to extract visual features $x_v = E_v(V)$, while a textual encoder $E_t$ processes the textual data to extract features $x_t = E_t(T)$. The extracted visual and textual representations are aligned in subsequent modules. Finally, the aligned features are fed into a classifier to produce the prediction $\hat{C}$, with model performance evaluated using specific metrics.

## 4 Method

This study proposes a Semantic-Space-Intervened Diffusive Alignment method (SeDA), which utilizes a modality-shared semantic space as an intermediary to enable the mapping from visual representations to textual representations, thereby effectively mitigating the heterogeneity between different modalities, as shown in Figure 2. SeDA consists of three primary modules: the Progressive Feature Interaction Network (PFIN), the Diffusion-controlled Semantic Learner (DSL), and the Diffusion-controlled Semantic Translator (DST). The DSL module operates in the early stage of the diffusion model, learning modality-independent semantic space from the original visual distribution. The DST module works in the later stage, projecting the semantic space to the textual features space. The PFIN module progressively introduces textual information by redesigning the diffusion model's network structure. Details of these modules are described below.

### 4.1 Progressive Feature Interaction Network (PFIN)

In the Progressive Feature Interaction Network module, we use cross-attention mechanisms and multimodal feature fusion to enable deep interaction between visual and textual information. This ensures effective integration and alignment of visual and textual features, enhancing cross-modal representation.

Unlike vanilla diffusion models [Ho *et al.*, 2020], which predict $\epsilon = \epsilon_\theta(x_i, i)$ using a UNet, We design the Feature Interaction and Reconstruction Network to predict $\tilde{x}_r = X_\theta(x_s^i, i, x_v)$ during the training phase of SeDA.

Specifically, given the visual features $x_v \in \mathbb{R}^{B \times d_k}$ and the textual features $x_s \in \mathbb{R}^{B \times d_k}$, where $B$ represents the batch size and $d_k$ denotes the feature dimension. In the forward process, Gaussian noise is added step-by-step to the data $x_s$, following a Markov process with a predefined variance schedule $\{\beta_i\}_{i=1}^T$:

$$q\left(x_s^i \mid x_s^0\right) = \mathcal{N}\left(x_s^i; \sqrt{\bar{\alpha}_i} x_s^0, (1 - \bar{\alpha}_i)\,\mathbf{I}\right) \tag{1}$$

where $x_s^0 \sim q(x)$, $\mathcal{N}(.)$ means a Gaussian distribution. $\alpha_i = 1 - \beta_i$ and $\bar{\alpha}_i = \prod_{j=0}^{i} \alpha_j$,

According to Equation 1, the noisy textual feature at step $i$ in the forward process is defined as $x_s^i = \sqrt{\bar{\alpha}_i} x_s + \sqrt{1 - \bar{\alpha}_i}\epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the added Gaussian noise. As $i \to +\infty$, $x_s^i$ undergoes a gradual convergence towards the standard Gaussian distribution.

The fused feature representation $F_t$ is generated through the cross-attention mechanism, and the process can be represented as:

$$\mathbf{F}_t = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where $x_s^i$ are used as the Query(Q), and the visual features $x_v$ are used as the Key(K) and Value(V).

SinusoidalPosEmb$(t)$ (sinusoidal position embedding function) is used to embed the time step $t$, obtaining the time feature vector $\mathbf{e}_t \in \mathbb{R}^{B\times d_k}$:

$$\mathbf{e}_t = \phi(\mathbf{W}_2 \cdot \phi(\mathbf{W}_1 \cdot \text{SinusoidalPosEmb}(t))) \tag{3}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices, and $\phi$ is the activation function.

Subsequently, we concatenate the output $\mathbf{F}_t$ from the attention module with the encoded diffusion time feature $\mathbf{e}_t$ as the input to the denoising decoder:

$$\mathbf{h}_0 = [\mathbf{F}_t; \mathbf{e}_t] \tag{4}$$

The denoising decoder is a multi-layer perceptron (MLP), which contains an intermediate layer with linear transformations and activation functions to encode the features, as well as a linear layer to compute the output distribution.

### 4.2 Diffusion-Controlled Semantic Learner (DSL)

Due to the heterogeneity between the visual and textual modalities, directly mapping from the visual modality to the textual modality poses challenges. To address this, we propose a bi-stage alignment strategy. In this module, we first map the visual features space to the semantic space by constraining the category centers of the visual features.

During this process, we construct the representation of $\tilde{x}_r$ by calculating the structural consistency loss between the interactive feature $\tilde{x}_r$ and the original input visual feature $x_v$.

The center of the visual features for category $c$ is denoted as $\mu_c$, and similarly, the center of the interactive features $\mu_c$ can be computed as:

$$\mu_c = \frac{1}{|x_v^c|}\sum_{x_v \in x_v^c} x_v, \quad \hat{\mu}_c = \frac{1}{|\tilde{x}_r^c|}\sum_{\tilde{x}_r \in \tilde{x}_r^c} \tilde{x}_r \tag{5}$$

Next, the structural consistency loss is defined as:

$$\mathcal{L}_{SC} = \|\hat{\mu}_c - \mu_c\|_2 + \sum_{x_v \in x_c^v, \tilde{x}_r \in \tilde{x}_c^r} \|\tilde{x}_r - x_v\|_1 \tag{6}$$

where the first term ensures the constraint of the feature centers, and the second term measures the offset between features.

Building upon cross-modal matching alignment, to enhance the model's performance in downstream visual classification tasks, we introduce a constraint cross entropy $\mathcal{L}_{CE}$ that combines visual prediction results $\hat{y} = softmax(\hat{x}_r^0)$ with real labels $y$:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{7}$$

The overall structural consistency constraint loss can be defined as:

$$\mathcal{L}_{SCC} = \alpha_1 \mathcal{L}_{SC} + \beta \mathcal{L}_{CE} \tag{8}$$

where $\alpha_1$ and $\beta$ are hyperparameters used to control the contributions of the classification loss and others, respectively.

### 4.3 Diffusion-Controlled Semantic Translator (DST)

In the later stage of the diffusion model, the DST module focuses on learning the underlying distribution of textual features while preserving semantic information. It facilitates the transformation from the semantic space to the textual features space.

To optimize the underlying data generation distribution, which is typically achieved by minimizing the variational lower bound (VLB) of the negative log-likelihood, We follow the DDPM [Ho *et al.*, 2020] setup and minimize the KL divergence between the two distributions $q(x_r^i|x_r^{i-1}, x_r^0)$ and $p_\theta(x_r^{i-1}|x_r^i)$:

$$\mathcal{L}_{vlb} = D_{KL}(q(x_r^i|x_r^{i-1}, x_r^0)||p_\theta(x_r^{i-1}|x_r^i)) \tag{9}$$

To simplify the optimization process, we reformulate it into a Mean-Squared Error (MSE) loss function as follows:

$$\mathcal{L}_{MSE} = \mathbb{E}_{x_s, x_s^i, x_v}[||x_s - X_\theta(x_s^i, i, x_v)||^2] \tag{10}$$

Similar to the DSCC module, we introduce the cross-entropy constraint to assist in the training of the DCT module:

$$\mathcal{L}_{CT} = \alpha_2 \mathcal{L}_{MSE} + \gamma \mathcal{L}_{CE} \tag{11}$$

where $\alpha_2$ and $\gamma$ are hyperparameters used to control the contributions of the MSE loss and the classification loss, respectively. The $\mathcal{L}_{CE}$ has been given in Equation7.

Finally, the diffusion model training process for this bi-stage can be expressed as follows:

$$\mathcal{L} = \sum_{i=0}^{T} \begin{cases} \alpha_1 \mathcal{L}_{\text{SC}} + \beta \mathcal{L}_{\text{CE}}, & \text{if } i \le t, \\ \alpha_2 \mathcal{L}_{\text{MSE}} + \gamma \mathcal{L}_{\text{CE}}, & \text{if } i > t. \end{cases} \tag{12}$$

where $T$ represents the diffusion model time step and $t$ represents the staged step.

### 4.4 Inference Phase of SeDA

In the inference process, the goal is to iteratively reconstruct the original data by optimizing the likelihood $p_\theta(x_r^0)$. The reverse process is defined by:

$$p_\theta\left(x_r^{i-1} \mid x_r^i\right) = \mathcal{N}\left(x_r^{i-1}; \mu_\theta\left(x_r^i, \tilde{x}_r\right), \tilde{\beta}_i \mathbf{I}\right) \tag{13}$$

where $\mu_\theta$ is the predicted mean, and $\tilde{\beta}_i = \frac{1-\bar{\alpha}_{i-1}}{1-\bar{\alpha}_i}\beta_i$ is the variance term.

According to [Song *et al.*, 2020], $\mu_\theta$ can be calculated using the predicted feature $\tilde{x}_r$ from the FIRN module:

$$\mu_\theta(x_r^i, \tilde{x}_r) = \frac{\sqrt{\bar{\alpha}_{i-1}}\beta_t}{1-\bar{\alpha}_i}\tilde{x}_r + \frac{\sqrt{\alpha_i}(1-\bar{\alpha}_{i-1})}{1-\bar{\alpha}_i}x_r^i \tag{14}$$

By performing a step-by-step reverse denoising operation:

$$\hat{x}_r^T \xrightarrow{p_\theta(\hat{x}_r^{T-1}|\hat{x}_r^T)} \hat{x}_r^{T-1} \xrightarrow{\cdots} \hat{x}_r^1 \xrightarrow{p_\theta(\hat{x}_r^0|\hat{x}_r^1)} \hat{x}_r^0 \tag{15}$$

the aligned feature $\hat{x}_r^0$ is finally obtained.

Subsequently, the aligned feature $\hat{x}_r^0$ is fed into a fully connected classifier to compute the predicted logits for the final classification.

| Method | Model | VIREO Food-172 | | NUS-WIDE | | | | MSRVTT | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Acc-1** | **Acc-5** | **Pre-1** | **Pre-5** | **Rec-1** | **Rec-5** | **Acc-1** | **Acc-5** |
| Visual Modal Backbone | ResNet-50 (CVPR'16) | 81.58 | 95.02 | 78.56 | 39.12 | 44.04 | 86.42 | 51.37 | 79.03 |
| | RepVGG (CVPR'21) | 83.47 | 96.03 | 79.71 | 39.44 | 44.82 | 85.58 | 50.96 | 77.12 |
| | RepMLPNet (CVPR'22) | 83.36 | 96.22 | 80.10 | 40.53 | 44.82 | 87.69 | 51.07 | 77.16 |
| | ViT-B/16 (ICLR'20) | 85.37 | 97.29 | 80.46 | 40.57 | 45.50 | 87.96 | 53.25 | 81.85 |
| | VanillaNet (NIPS'24) | 84.51 | 96.04 | 80.12 | 39.46 | 45.65 | 85.73 | 52.47 | 80.02 |
| Alignment Framework | ATNet (MM'19) | 85.67 | 96.81 | 80.78 | 39.89 | 45.59 | 86.55 | 54.45 | 82.68 |
| | CLIP (PMLR'21) | 85.56 | 96.98 | 81.64 | 40.87 | 46.25 | 88.78 | 52.89 | 81.66 |
| | TEAM (MM'22) | 87.70 | 97.85 | 81.98 | 40.80 | 46.48 | 88.54 | 55.08 | 84.12 |
| | ITA (CVPR'22) | 87.82 | 97.89 | 82.65 | 41.40 | 46.99 | 89.13 | 55.05 | 84.05 |
| | SDM (CVPR'23) | 87.63 | 97.78 | 82.64 | 41.20 | 47.03 | 89.31 | 54.82 | **84.28** |
| | MM-PKD (CVPR'23) | 87.89 | 96.96 | 81.76 | 41.31 | 47.24 | 89.11 | 54.77 | 82.45 |
| | C2KD (CVPR'24) | 87.83 | 98.06 | 82.77 | 41.25 | 47.20 | 89.35 | 55.15 | 82.21 |
| | MGCC (AAAI'24) | 87.80 | 97.87 | 82.09 | 41.05 | 46.69 | 89.14 | 54.88 | 83.67 |
| | MoMKE (MM'24) | 87.97 | 97.09 | 81.72 | 41.11 | 46.50 | 89.14 | 54.57 | 82.72 |
| | SeDA$_{RN50}$ | 86.01 | 96.97 | 81.60 | 40.52 | 46.25 | 87.96 | 54.68 | 81.84 |
| | SeDA$_{ViT16}$ | **89.19** | **98.07** | **83.46** | **41.60** | **47.72** | **90.21** | **57.09** | 83.91 |

Table 1: Performance comparison of algorithms on VIREO Food-172, NUS-WIDE and MSRVTT datasets. Metrics are Top-1/Top-5 Accuracy (Acc), Precision (Pre), and Recall (Rec). The best performance of each indicator has been highlighted in bold

# 5 Experiment

## 5.1 Experiment Setting

**Datasets**

To assess the effectiveness and generality of SeDA, we conducted experiments on image and video classification tasks across three datasets. Details are provided below.

- **VIREO Food-172[Chen and Ngo, 2016]:** A single-label dataset with 110,241 food images in 172 categories and an average of three text descriptions per image. It includes 66,071 training and 33,154 test images.
- **NUS-WIDE[Chua *et al.*, 2009]:** A multi-label dataset of 203,598 images (after filtering) in 81 categories, with textual tags from a 1000-word vocabulary. It has 121,962 training and 81,636 test images.
- **MSRVTT[Xu *et al.*, 2016]:** A video dataset with 10,000 YouTube clips and 200,000 captions. We used 7,010 videos for training and 2,990 for testing.

**Evaluation Metric**

For the VIREO Food-172 and MSRVTT datasets of the single-class prediction task, we use the accuracy of Top-1 and -5 following [Meng *et al.*, 2019]. For the multi-label dataset NUS-WIDE, we calculate Top-1 and -5 overall precision and recall following [Wang *et al.*, 2017; Gong *et al.*, 2013].

**Implementation Details**

In this experiment, we chose Adam as the optimizer for the model, with a weight decay of 1e-4. The learning rate for all neural networks was set between 1e-4 and 5e-5. The learning rate decayed to half of its original value every four training epochs. For the loss weights mentioned in the training strategy, we selected $\alpha_1$ and $\alpha_2$ between 0.1 and 2.0, the time step $T$ between 900 and 1500, the staged step $t$ between 0 and 500, while $\beta$ and $\gamma$ were chosen from [0.5, 1.0, 1.5, 2.0]. Our experiments were conducted on a single NVIDIA Tesla V100 GPU, using PyTorch 1.10.2, and the batch size is 64.

## 5.2 Performance Comparison

We conducted a comprehensive comparison involving 5 visual modal backbones and 9 alignment frameworks, with results summarized in Table 1. The visual modal backbones were implemented based on the methods outlined in their respective papers, including ResNet-50 [He *et al.*, 2016], RepVGG [Ding *et al.*, 2021], RepMLPNet [Ding *et al.*, 2022], ViT-B/16 [Dosovitskiy *et al.*, 2020], and VanillaNet [Chen *et al.*, 2024]. For the alignment framework, ViT-B/16 was used to encode the visual channel, while BERT [Devlin, 2018] was employed for the textual channel. In addition, we validated the effectiveness of our method on ResNet-50. The cross-modal alignment methods were adapted from their original designs to fit the LUPI task, including ATNet [Meng *et al.*, 2019], CLIP [Radford *et al.*, 2021], TEAM [Xie *et al.*, 2022], ITA [Wang *et al.*, 2022], SDM [Jiang and Ye, 2023], MM-PKD [Aslam *et al.*, 2023], C2KD [Huo *et al.*, 2024], MGCC [Wu *et al.*, 2024], and MoMKE [Xu *et al.*, 2024]. A fully connected single-layer network served as the classifier.

- **The proposed method, SeDA, outperforms other algorithms across three datasets.** This is attributed to our method's ability to learn the underlying distribution of textual features, helping alleviate semantic confusion between categories.
- **SeDA is a general framework that can combine various visual backbones,** such as ViT-B/16 and ResNet-50, to bring them performance gains, which showcases its model-agnostic capability.
- **Models in the Alignment Framework generally outperform the Visual Modal Backbone.** This can be at-

| Method | VIREO Food-172 | | NUS-WIDE | | | |
|---|---|---|---|---|---|---|
| | Acc-1 | Acc-5 | Pre-1 | Pre-5 | Rec-1 | Rec-5 |
| Base | 85.37 | 97.29 | 80.46 | 40.57 | 45.50 | 87.96 |
| +T | 88.70 | 96.93 | 82.63 | 41.23 | 47.19 | 89.37 |
| +T+I | 89.12 | 97.01 | 82.74 | 41.20 | 47.26 | 89.48 |
| +T+I+L | **89.19** | **98.07** | **83.46** | **41.60** | **47.72** | **90.21** |

Table 2: Results of ablation study. The evaluation indexes are the same as those in Table 1. The best performance is marked in bold.

tributed to the incorporation of more discriminative textual information, which effectively optimizes the visual representation.

- **SeDA shows a more significant improvement in Top-1 accuracy than in Top-5.** This is mainly because it focuses on learning the distribution of textual features, aiming to accurately provide the most discriminative first label rather than focusing on improving the precision in retrieval ranking.

## 5.3 Ablation Study

This section examines the performance of different modules based on ViT-B/16 in SeDA, with the results presented in Table 2.

- **On the VIREO Food-172 dataset, Diffusion-controlled Semantic Translator (+T) module played a key role in improving Acc-1, achieving a 3.7% increase compared to the baseline method.** However, there was a slight decrease in Acc-5.
- **Adding the Progressive Feature Interaction Network (+I) module led to improvements across both datasets.** This indicates that it (+I) further enhances feature interaction capabilities.
- **Diffusion-controlled Semantic Learner (+L) module ensured a certain improvement in the Top-1 metric while achieving a more significant enhancement in the Top-5 metric.** Notably, on the NUS-WIDE dataset, Rec-5 improved by 2.6% compared to the baseline method.

## 5.4 In-depth Analysis

### Robustness of SeDA on Hyperparameters

This section evaluates the robustness of SeDA in different hyperparameters. We select the weight parameter $\gamma$, time step $T$ and stages step $t$ from $\{0.5,1.0,1.5,2.0\}$, $\{900,1200,1500,1800\}$ and $\{0,300,500\}$. **SeDA is largely insensitive to changes in hyperparameters, demonstrating strong robustness in hyperparameter selection.** For $\gamma$, the model performs best when $\gamma = 1.5$ This is because lower $\gamma$ values tend to rely overly on specific features at the expense of class-level information, whereas higher $\gamma$ values place excessive emphasis on class information, disrupting the learning of the feature distribution space. Furthermore, the model achieves its best performance when the diffusion model time step $T$ is set to 1500, as smaller $T$ fails to fully learn useful information for transferring the visual modality to the text modality, while larger $T$ may introduce additional noise. Notably, **two-stage learning is beneficial, but excessively long staged step $t$ in the second stage may damage the structural information of the original features, leading to a slight drop in performance.**

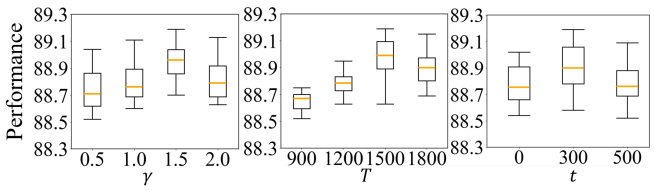


Figure 3: The impact of hyperparameters on performance. The weight parameter $\gamma$, the time step $T$ and the staged step $t$ are turned from $\{0.5,1.0,1.5,2.0\}$, $\{900,1200,1500,1800\}$, $\{0,300,500\}$ on VIREO Food-172, respectively.

### The Effect of Different Methods in Each Module

To further investigate the impact of different methods in the DSL, DST, and PFIN modules on model performance, we used ViT-B/16 as the baseline and conducted experiments on the VIREO Food-172 dataset, as shown in Tabel 3. Four aspects were analyzed: training strategies, diffusion model network structures, modal interaction methods, and stage training methods. Fixing text features limits flexibility, while joint optimization improves cross-modal understanding. UNet is better suited for pixel-level learning, while MLP, which outperforms UNet, is more effective for feature-level optimization. Cross Attention proved superior for modal interaction, effectively capturing correlations. Bi-stage framework effectively integrates semantic information from visual and text features, enhancing alignment and classification capabilities.

## 5.5 Case Study

### Performance of Visual to Textual Feature Distribution Transformation

We randomly selected a category to compare the textual and visual features of the ViT-B/16, features aligned using traditional L2-norm alignment (as shown in Figure 4(a)), and features reconstructed by our method during the diffusion model sampling process across different time steps (Figure 4 (b)). The feature distribution produced by traditional alignment methods demonstrates minimal changes, with significant discrepancies still evident between the aligned visual and textual features. This suggests that traditional methods fail to bridge the gap between the two modalities fully. In contrast, the features reconstructed using our method initially capture the semantic characteristics of visual features, gradually moving toward textual features as the sampling process progresses. This dynamic migration process ensures that visual features are effectively transformed into a semantic space consistent with textual features. Furthermore, our method not only achieves accurate mapping between the two modalities but also enhances the expressiveness of the textual features by preserving and utilizing semantic information.

| Strategy | Method | Acc-1 | Acc-5 |
|---|---|---|---|
| **Base** | ViT-B/16 | 85.37 | 97.29 |
| **Different Training Strategies** | Freezing the visual and the textual backbone model | 87.27 | 93.45 |
| | Training the visual backbone, freezing the textual backbone | 88.67 | 94.84 |
| | Training the visual and the textual backbone model | 88.70 | 96.93 |
| **Different Diffusion Model Network Structures** | Using UNet as the diffusion model network | 86.98 | 92.22 |
| | Using MLP as the diffusion model network | 87.27 | 93.45 |
| **Different Modal Interaction Methods** | Using Concat for modal interaction | 88.70 | 96.93 |
| | Using Self Attention for modal interaction | 88.21 | 96.46 |
| | Using Multimodal Transformer for modal interaction | 87.62 | 95.44 |
| | Using Cross Attention for modal interaction | 89.12 | 97.01 |
| **Different Stage Training Methods** | Using a single-stage method with only the DSL module | 88.97 | **98.13** |
| | Using a single-stage method with only the DST module | 89.12 | 97.01 |
| | Bi-stage method with DSL module first and DST module second | **89.19** | 98.07 |

Table 3: Results from the combined experiment using different strategies and constraint functions on the VIREO Food-172 dataset.

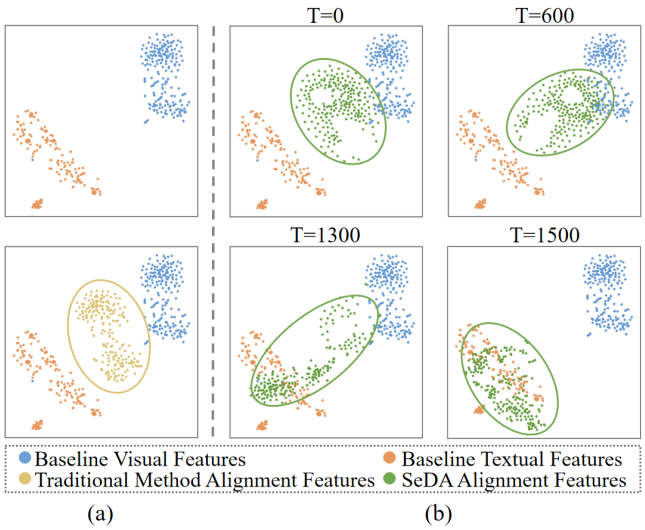


Figure 4: T-SNE visualization of data distribution before and after alignment for a randomly selected category.

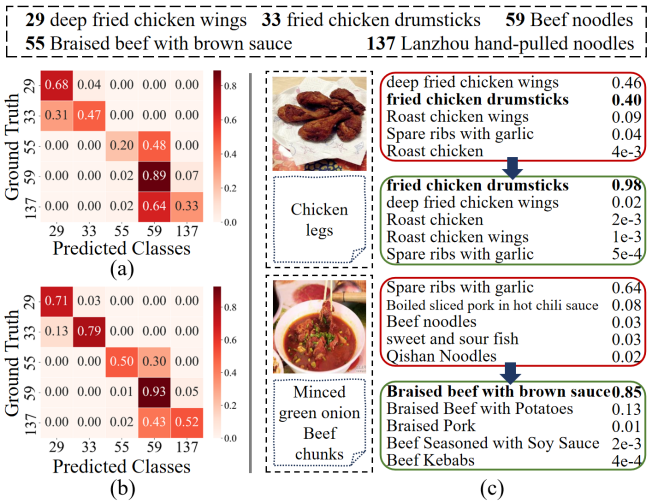


Figure 5: Comparison of ViT-B/16 and SeDA on confusion matrix and randomly selected samples with Top-5 confidence scores. Red boxes represent baseline results, green boxes represent SeDA results, and ground-truth labels are highlighted in bold.

### Effectiveness of Semantic Disambiguation

We selected five categories with severe confusion in the Baseline model, as shown in Figure 5(a). The corresponding confusion matrix for SeDA is shown in Figure 5(b). The results clearly demonstrate that confusion between semantically similar categories is significantly reduced, leading to an effective improvement in classification performance. Additionally, Figure 5(c) presents a detailed analysis of the prediction results for specific samples. For example, in the case of the "deep fried chicken wings" sample, the model accurately predicted the correct category with high confidence after incorporating textual information. This indicates that our method effectively leverages semantic-related information to enhance the distinction between easily confused categories. On the other hand, for samples with substantial background noise (e.g., "Braised beef with brown sauce"), the Baseline model tends to misclassify them into irrelevant categories such as "noodles" or "garlic" due to noise factors like "chopsticks" and "onion" in the original image. However, SeDA successfully extracts critical features, accurately narrowing the predicted category to the semantic domain related to "beef," thereby significantly mitigating classification errors caused by semantic ambiguity.

## 6 Conclusion

To address the heterogeneity between modalities, this paper transfers the multi-step denoising process of diffusion models to the cross-modal alignment of visual representations, proposing a semantic-space-intervened diffusive alignment method (SeDA). Using semantic space as an intermediary, a bi-stage diffusion model alignment is designed: the DSL module first captures the semantic information of visual features, and the DST module gradually translates semantic features to textual features. This method effectively projects from visual representations to textual representations. Despite SeDA alleviating the heterogeneity between different modalities, there remains room for improvement in fine-grained alignment. Future work will explore fine-grained modeling with diffusion to further improve alignment accuracy, especially in complex scenarios and datasets.

## Acknowledgments

This work is supported in part by the Shandong Province Excellent Young Scientists Fund Program (Overseas) (Grant no. 2022HWYQ-048).

## References

[Aslam *et al.*, 2023] Muhammad Haseeb Aslam, Muhammad Osama Zeeshan, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, and Eric Granger. Privileged knowledge distillation for dimensional emotion recognition in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3338–3347, 2023.

[Aslam *et al.*, 2024] Muhammad Haseeb Aslam, Muhammad Osama Zeeshan, Soufiane Belharbi, Marco Pedersoli, Alessandro Lameiras Koerich, Simon Bacon, and Eric Granger. Distilling privileged multimodal information for expression recognition using optimal transport. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024.

[Baltrušaitis *et al.*, 2019] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[Chen and Ngo, 2016] Jingjing Chen and Chong-Wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 32–41, 2016.

[Chen *et al.*, 2023] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19830–19843, October 2023.

[Chen *et al.*, 2024] Hanting Chen, Yunhe Wang, Jianyuan Guo, and Dacheng Tao. Vanillanet: the power of minimalism in deep learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[Dang *et al.*, 2023] Jisheng Dang, Huicheng Zheng, Jinming Lai, Xu Yan, and Yulan Guo. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *IEEE Transactions on Image Processing*, 32:3924–3938, 2023.

[Dang *et al.*, 2024a] Jisheng Dang, Huicheng Zheng, Bimei Wang, Longguang Wang, and Yulan Guo. Temporo-spatial parallel sparse memory networks for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):17291–17304, 2024.

[Dang *et al.*, 2024b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, and Yulan Guo. Unified spatio-temporal dynamic routing for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(5):4512–4526, 2024.

[Dang *et al.*, 2024c] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, and Yulan Guo. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *IEEE Transactions on Image Processing*, 33:4853–4866, 2024.

[Dang *et al.*, 2025] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, Qingyong Hu, and Yulan Guo. Adaptive sparse memory networks for efficient and robust video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3820–3833, 2025.

[Deng *et al.*, 2025] Bowen Deng, Tong Wang, Lele Fu, Sheng Huang, Chuan Chen, and Tao Zhang. THESAURUS: contrastive graph clustering by swapping fused gromov-wasserstein couplings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16199–16207, 2025.

[Devlin, 2018] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Ding *et al.*, 2021] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, pages 13733–13742, 2021.

[Ding *et al.*, 2022] Xiaohan Ding, Honghao Chen, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Repmlpnet: Hierarchical vision mlp with re-parameterized locality. In *CVPR*, pages 578–587, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fu *et al.*, 2025] Lele Fu, Sheng Huang, Yanyi Lai, Tianchi Liao, Chuanfu Zhang, and Chuan Chen. Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16648–16656, 2025.

[Gong *et al.*, 2013] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-

nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Ho *et al.*, 2022] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[Huo *et al.*, 2024] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015, 2024.

[Jiang and Ye, 2023] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023.

[Jin *et al.*, 2023] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2470–2481, 2023.

[Lee *et al.*, 2019] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10285–10295, 2019.

[Li *et al.*, 2024] Xiang Li, Lei Meng, Lei Wu, Manyi Li, and Xiangxu Meng. Dreamfont3d: personalized text-to-3d artistic font generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[Meng *et al.*, 2019] Lei Meng, Long Chen, Xun Yang, Dacheng Tao, Hanwang Zhang, Chunyan Miao, and Tat-Seng Chua. Learning using privileged information for food recognition. In *ACM MM*, pages 557–565, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[Sun *et al.*, 2016] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[Wang *et al.*, 2017] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017.

[Wang *et al.*, 2022] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022.

[Wang *et al.*, 2023] Changshuo Wang, Lei Wu, Xiaole Liu, Xiang Li, Lei Meng, and Xiangxu Meng. Anything to glyph: artistic font synthesis via text-to-image diffusion model. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[Wang *et al.*, 2024] Yuqing Wang, Lei Meng, Haokai Ma, Yuqing Wang, Haibei Huang, and Xiangxu Meng. Modeling event-level causal representation for video classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3936–3944, 2024.

[Wu *et al.*, 2023] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023.

[Wu *et al.*, 2024] Xinyi Wu, Wentao Ma, Dan Guo, Tongqing Zhou, Shan Zhao, and Zhiping Cai. Text-based occluded person re-identification via multi-granularity contrastive consistency learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6162–6170, 2024.

[Xie *et al.*, 2022] Chen-Wei Xie, Jianmin Wu, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Token embeddings alignment for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4555–4563, 2022.

[Xu *et al.*, 2016] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[Xu *et al.*, 2024] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446, 2024.

[Yang *et al.*, 2025] Yimeng Yang, Haokai Ma, Lei Meng, Shuo Xu, Ruobing Xie, and Xiangxu Meng. Curriculum conditioned diffusion for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13035–13043, 2025.