Preserving Modality Structure Improves Multi-Modal Learning

Sirnam Swetha¹, Mamshad Nayeem Rizve¹, Nina Shvetsova²,³, Hilde Kuehne²,³,⁴, Mubarak Shah¹

¹ CRCV, University of Central Florida, ²Goethe University Frankfurt Germany, ³University of Bonn Germany, ⁴MIT-IBM Watson AI Lab

Abstract

Self-supervised learning on large-scale multi-modal datasets allows learning semantically meaningful embeddings in a joint multi-modal representation space without relying on human annotations. These joint embeddings enable zero-shot cross-modal tasks like retrieval and classification. However, these methods often struggle to generalize well on out-of-domain data as they ignore the semantic structure present in modality-specific embeddings. In this context, we propose a novel Semantic-Structure-Preserving Consistency approach to improve generalizability by preserving the modality-specific relationships in the joint embedding space. To capture modality-specific semantic relationships between samples, we propose to learn multiple anchors and represent the multifaceted relationship between samples with respect to their relationship with these anchors. To assign multiple anchors to each sample, we propose a novel Multi-Assignment Sinkhorn-Knopp algorithm. Our experimentation demonstrates that our proposed approach learns semantically meaningful anchors in a self-supervised manner. Furthermore, our evaluation on MSR-VTT and YouCook2 datasets demonstrates that our proposed multi-anchor assignment based solution achieves state-of-the-art performance and generalizes to both in-and out-of-domain datasets. Code: https://github.com/Swetha5/Multi_Sinkhorn_Knopp

1. Introduction

Humans often rely on multiple sensory inputs to have a better understanding of everyday events. Most commonly, we utilize vision, audio, and language to perceive an event as they provide complementary information for robust reasoning. The closest approximation of this setup is video data as it provides both visual and audio information along with a text description as a caption. Recently, researchers have started to explore learning meaningful representations by leveraging multiple modalities to train efficient models at scale [5, 11, 42]. Such systems focus on representation learning that either improves features for each modality separately [5] or learns a joint multi-modal embedding [11, 42] space that enables various zero-shot tasks like retrieval or classification. However, given the inherent differences across the modalities, it is challenging to learn effective joint embeddings. Furthermore, the real-world data presents additional challenges like misalignment between modalities, leading to weak supervision.

Current pre-training approaches in this area usually employ a contrastive objective [33] to learn the joint embeddings that pulls the cross-modal embeddings of a sample from the same temporal instance closer and pushes embeddings of other samples farther. Despite promising performances, these methods struggle with generalizability. This is particularly evident in previous approaches trained on HT100M [11, 42], which do well on the closely related downstream dataset YouCook2 but struggle to improve on the MSR-VTT dataset, which exhibits a relatively larger domain shift with respect to HT100M [42]. This is due to the contrastive objective’s emphasis on strict alignment between modalities in the joint embedding space while ignoring the inherent weak alignment between different modalities [45], as well as the underlying semantic structure across samples [43, 50]. Recent works have tackled these issues, either by using joint multi-modal clustering [11] to preserve the semantic structure in the joint embedding space or by incorporating a reconstruction objective [11, 25] to retain modality-specific features in the joint embedding space, allowing for weak multi-modal alignment. However, the usual reconstruction objective trivially tries to retain most modality-specific features in the joint space, thus preventing the learning of optimal features for cross-modal tasks. And the multi-modal clustering approaches perform hard-clustering making it less flexible. Therefore, the limitations of the contrastive objective cannot be adequately addressed even after combining these independent objectives.

To address this, we propose a semantic-structure-
preserving consistency loss (SSPC) to only retain information that is beneficial for both cross-modal embedding learning and retaining modality-specific semantic structure. In particular, for SSPC loss we consider each sample (e.g., a video clip) to be composed of multiple concepts: scene or objects involved in the downstream task. Therefore, the relationship between samples is multifaceted, representing both shared and unique concepts across samples. To capture this multifaceted relationship in a flexible manner, we propose to learn anchors (latent codes) and model the relationship between samples with respect to their relationships with these anchors. Therefore, these anchors act as a proxy to represent the modality-specific relationships between samples (semantic structure) which can be preserved using the proposed SSPC loss. Since we have no supervision to learn these anchors, we formulate this anchor learning problem as a many-to-many assignment problem, as modeling this multifaceted relationship simultaneously involves assigning multiple anchors to one sample and multiple samples to one anchor. Although there is a vast literature on solving the many-to-one assignment problem [9, 5, 38, 48], there is no efficient way to solve this many-to-many assignment problem.

To this end, we propose a novel Multi-Assignment Sinkhorn-Knopp (Multi-SK) algorithm that iteratively optimizes the many-to-many anchor assignments for both the modality-specific embeddings (in input space) and modality-agnostic multi-modal embeddings (in joint embedding space). To allow for weak alignment between modalities, we select the dominant anchors for each sample to represent the relationship between different samples. Our proposed SSPC loss enforces consistency between the dominant anchor assignments at the input and joint embedding spaces to preserve the modality-specific semantic structure. To demonstrate the effectiveness of our proposed solution, we train our model on HT100M dataset and test on 6 zero-shot tasks on multiple downstream datasets and observe that our approach leads to state-of-the-art results in all settings.

In summary, we make the following contributions: (i) We propose a flexible modality-specific semantic-structure-preserving approach to improve the generalizability of cross-modal features. (ii) We introduce Multi-Assignment Sinkhorn-Knopp, a novel algorithm to enable multiple assignments for flexible sample relationship modeling. (iii) Our proposed method outperforms the current state-of-the-art for multi-modal self-supervised representation learning on both in- and out-of-domain datasets.

2. Related Work
2.1. Multi-Modal Learning

With the availability of large-scale multi-modal datasets [31, 41, 7], multi-modal learning research has received a lot of attention. It comprises of vision-language learning [35, 51], vision-audio learning [2, 4, 6, 12, 46], video-audio-language learning [39, 11, 42], zero-shot learning [21, 28], cross-modal generation [37, 53, 27] and multi-modal multi-task learning [22]. Miech et al. [31] proposed a large-scale multi-modal dataset consisting of video, audio and text by collecting instructional videos from YouTube without requiring any human annotations. Note that the text is generated from audio using Automatic Speech Recognition (ASR) and has noisy alignment between the text and video. They also proposed a multi-modal system to demonstrate the potential for learning video-text embedding via contrastive loss. To handle the noise in the dataset, Amrani et al. [3] proposed a noise estimation for multi-modal data via multi-modal density estimation. A noise-contrastive estimation approach in a multi-instance learning framework was proposed by Miech et al. [29]. XDC [2] performs clustering on audio-video for learning better features for each modality separately. These works utilize only two modalities for multi-modal learning, while others have explored utilizing audio, video, and text together for multi-modal learning. Multi-Modal versatile networks [1] was proposed to learn different embedding spaces for each combination of modalities. AVLNet [39] proposed to learn a shared embedding that maps all modalities to a single joint embedding space. Following this, MCN [11] proposed to perform joint clustering and reconstruction to learn joint embedding space. Note that, [11] performs multi-modal K-means clustering to learn hard semantic clusters. Unlike strict assignment in [11], we propose flexible learning with multiple assignments and separately for each modality. More recently, EAO [42] utilizes transformers and combinatorial fusion of modalities to learn the joint embedding with contrastive loss.

Most of these works, utilize contrastive or clustering loss over fused multi-modal representation to learn the joint embedding space. By doing so, these models do not retain the modality-specific semantic structuring between samples encoded by the pre-trained modality specific backbones, hurting the generalization ability of the model. Additionally, recent works have reported that large-scale contrastive multi-modal models (e.g., CLIP [36]) are somewhat robust to distributional shifts mainly due to diverse large-scale training data and prompt-engineering [15]. Therefore, our work focuses on making the pre-training objective robust to distributional shifts. In this context, we propose a novel approach to preserve the modality-specific semantic relationships in the joint embedding space by modeling the relationship between samples w.r.t learnable anchors. To enable flexible relationship modeling between samples, we learn multiple anchor assignments per sample, where anchors shared across samples model the commonality between them, and the distinct anchors between samples highlight the uniqueness of the samples.
2.2. Sinkhorn-Knopp

Recently, Sinkhorn-Knopp algorithm [44] has drawn huge attention because of its effectiveness in solving optimal-transport problems [23, 8]. Specifically, [13] proposed an entropic relaxation of the optimal transport problem which can be efficiently solved using Sinkhorn’s matrix scaling algorithm. Many following works have since successfully utilized the Sinkhorn-Knopp algorithm to solve different label assignment problems framed as an optimal transport problems. For instance, SeLa [48] cast the unsupervised clustering problem as a pseudo-label assignment problem and used the Sinkhorn-Knopp algorithm to solve it. SeLaVi [5] extended this idea to self-supervised representation learning for multi-modal data where the cluster assignments between different modalities are swapped to encourage modality invariant representation learning. Similarly, SwAV [9] used the Sinkhorn-Knopp algorithm for self-supervised representation learning and proposed to swap the pseudo-labels for differently augmented versions of a sample and use soft assignments instead of hard pseudo-labels. In contrast to these works, SuperGlue [40] used the Sinkhorn-Knopp algorithm to solve the correspondence problem between two sets of local features. Moreover, Sinkhorn-Knopp has been used in detection problems [17], where it was used to match the anchors with ground truths. Recently, UNO [16], and TRSSL [38] have successfully used the Sinkhorn-Knopp algorithm in solving novel class discovery and open-world semi-supervised learning problems, respectively. One key limitation of the traditional Sinkhorn-Knopp algorithm is that it cannot be directly utilized to compute multiple assignments necessary to perform multi-anchor based learning i.e. many-to-many assignments.

Some prior works [18, 26] have attempted to solve many-to-many assignments in indirect ways. While authors of [18] use an intermediate graph to match groups of vertices from source to target graph, which can only perform group-to-group assignments and is inadequate for our problem that requires true many-to-many matching. [26] modifies the Sinkhorn-Knopp row and column constraints to obtain many-to-many assignment to model dense correspondences, however, we have found this approach to be inferior in solving the multiple anchor assignment problem. The modified Sinkhorn-Knopp constraints approach yields suboptimal results, as discussed in Sec. 4.4. To address these
limitations, we propose a novel algorithm Multi-SK, which outperforms the modified Sinkhorn-Knopp constraints approach to get true many-to-many assignments.

3. Method

Given a set of multi-modal inputs \( \{t^{(i)}, v^{(i)}, a^{(i)}\}_{i=1}^N \) from \( N \) video clips, we learn modality-specific projection functions \( f_t, f_v, f_a \), that transform \( t, v, a \) into a \( d \)-dimensional joint embedding space, \( \mathbb{R}^d \), to obtain \( \tilde{t}, \tilde{v}, \tilde{a} \) respectively. Our goal is to optimize the parameters of \( f_t, f_v, f_a \) in such a way that they maintain the semantic structure amongst samples from a particular modality in the joint embedding space, as discussed in Sec 3.1, and simultaneously brings the semantically related cross-modal inputs closer. In the following, first, we formulate our approach to modeling the relationship between samples using anchors in Sec 3.1, then in Sec 3.2, we discuss our novel Multi-SK algorithm to learn these anchors for representing sample relationships, and finally, we present the overall training objective to train the model in Sec 3.3.

### 3.1. Modeling Sample Relationships with Anchors

Our work aims to preserve the relationship between samples \( i.e., x^{(i)}, x^{(j)} \) from a particular modality, to better generalize on the unseen data. To this end, we propose to model the relationship between samples using anchors, where the similarity of each sample w.r.t. anchors encodes the semantic structure of the feature space. Unlike clustering approaches, which involve hard-assignment to a specific cluster, our approach offers flexibility \( i.e., \) there can be shared and unique anchors across samples that define the relationship between them. To be particular, for each sample, we learn \( K \) anchors, and the similarity of assignments over these anchors represents the relationship (which we want to preserve) between samples from a particular modality.

As the pre-trained features extracted from modality-specific backbones encode the semantic structure between samples within that particular modality, preserving such modality-specific relationships in the joint embedding space would boost the generalizability of the model. To achieve this, we define two sets of \( K \) learnable anchors \( z = \{z^{(i)}\}_{i=1}^K \), and \( \hat{z} = \{\hat{z}^{(i)}\}_{i=1}^K \) to model the sample relationships before and after projecting them to the joint embedding space respectively. We repeat this for all the modalities, \( t, v, a \). We propose to preserve the modality-specific semantic structure at the joint embedding space by enforcing consistency in anchor assignments before and after performing feature projections (Eq. 3).

We still have one remaining challenge, \( i.e., \) how to learn/discover these anchors in an unsupervised manner. To this end, we cast the anchor discovery as a label assignment task with a uniform prior, \( i.e., \) each anchor will have an equal number of sample assignments. Additionally, to encourage flexible modeling, we enforce multiple anchor assignments per sample. However, it is difficult to estimate the exact number of anchors for each sample without additional prior information, thus we select the top \( K' \) anchors for each sample to effectively model the sample relationships. Even though this optimization task may seem like a difficult combinatorial problem, we model this combinatorial optimization task as an optimal transport problem. Following recent works [5, 9], one might assume the Sinkhorn-Knopp algorithm is a natural choice to solve this problem, however, the vanilla Sinkhorn-Knopp algorithm cannot handle multiple assignments per sample \( i.e., \) many-to-many assignment. To address this limitation, we propose Multi-Assignment Sinkhorn-Knopp algorithm, presented in the following.

### 3.2. Multi-Assignment Sinkhorn-Knopp

Given a sample matrix \( B \) s.t. \( B \in \mathbb{R}^{N \times d} \), representing \( N \) samples and an anchor matrix \( Z \) s.t. \( Z \in \mathbb{R}^{K \times d} \), of \( K \) anchor vectors, we obtain the similarity matrix \( S \) s.t. \( S = BZ^\top \) and \( S \in \mathbb{R}^{N \times K} \), where, \( S_{ij} \) is the probability of assigning \( j^{th} \) anchor to the \( i^{th} \) sample vector. The goal is to find an anchor assignment matrix, \( Q \), such that it satisfies the following constraints: (i) a sample should be assigned to exactly \( K' \) anchors, to learn top \( K' \) anchor assignments per sample. (ii) anchor assignments must be equally partitioned \( i.e, \) each anchor must be selected exactly \( N \times K' / K \) times, for uniform anchor assignment.

To obtain such multiple anchor assignments per sample, we propose to create a substitute 3D assignment matrix \( Q' \) s.t. \( Q' \in \mathbb{R}^{K \times N \times K} \). We also generate a 3D similarity matrix \( S' \) from the 2D similarity matrix \( S \) by introducing \( K \) channels (depth dimension) to get \( K \times N \times K \) matrix such that each channel is a scaled matrix of \( S \) with a predefined ranking between the channels enabling top \( K' \) anchor selection. Since we are only interested in selecting the top \( K' \) anchors, we set the first \( K' \) channels of \( S' \) to be the same as \( S \). The remaining \( K - K' \) channels are set to \( \mu S \). Here, \( \mu \) is a damping factor s.t. \( 0 < \mu < 1 \), to help select the top \( K' \) anchors for each sample. We discuss alternate designs for \( S' \) generation in the Supplementary Sec. 2.

The optimization objective of Multi-Assignment Sinkhorn-Knopp is to find an assignment matrix \( Q' \) such that it satisfies our multi-anchor assignment constraints while maximizing similarity with the initial assignment/similarity matrix, \( S' \). This optimization problem is defined as:

\[
Q^* : \max_{Q'} < Q', S' > + \epsilon H(Q')
\]

\[
H(Q') = -\sum Q'_{ijk} \log Q'_{ijk}
\]

\( Q^* \) needs to satisfy the following constraints to be a valid solution for our multi-anchor assignment problem.

- **Q' Row Constraint**: Within a channel, the sum of all el-
ements in a particular row must be equal to one. This is because we only want one anchor assignment for a sample in a particular channel. \( \forall i, k \sum_j Q^*_{ijk} = 1 \)

- **Q* Column Constraint**: In a channel, the sum of all elements in a column should be equal to \( N/K \). This constraint enforces equal partitioning of anchor assignments. \( \forall j, k \sum_i Q^*_{ijk} = N/K \).

- **Q* Depth constraint**: Depth-wise sum should be equal to one for every sample and anchor combination. This constraint prevents selecting the same anchor across different channels. \( \forall i, \sum_k Q^*_{ijk} = 1 \).

The traditional Sinkhorn-Knopp method uses an iterative matrix scaling algorithm that scales the rows and columns alternatively (with desired constraints) till the desired assignment matrix is obtained. We employ a similar scheme and extend the iterative scaling to the depth dimension for estimating \( Q^* \). We iteratively scale the rows, columns, and channels (depth dimension) till convergence. The final 2D assignment matrix \( Q \) is computed by performing a depth-wise sum on the top \( K \) channels. We provide Pytorch-style pseudo-code for our Multi-Assignment Sinkhorn-Knopp algorithm in the Supplementary Sec. 3.

### 3.3. Training Objective

#### Semantic Structure Preserving Consistency Loss.

To preserve the semantic structure of each modality, \( t, v, a \), we apply consistency loss to enforce similar anchor assignments between input and joint embedding space. As the cross-modal contrastive loss in the joint embedding space tries to bring different modalities together, features in the joint embedding space from a particular modality should preserve the common anchors that exist in corresponding features from the other modalities. Therefore, we also apply cross-modal anchor consistency across all modalities as shown in Eq. 3. Since we are dealing with 3 input modalities, this cross-modal consistency results in 9 consistency constraints.

Let’s denote \( L(t, \hat{v}, \hat{z}, \hat{\tilde{z}}) \) as the consistency loss between text anchor assignment in the input space and the textual anchor assignment of the corresponding video features at the joint embedding space as shown below in Eq. 2:

\[
L(t, \hat{v}, \hat{z}, \hat{\tilde{z}}) = \alpha_t, g(\text{sim}(t, \hat{z}_i), M(\text{sim}(\hat{v}, \hat{\tilde{z}}))) \\
+ \beta_t, g(\text{sim}(\hat{v}, \hat{\tilde{z}}), M(\text{sim}(t, \hat{z}_i)))
\]

Here, \( z_i, \hat{z}_i \) respectively represent input and output learnable anchor vectors for the text modality. \( g(.,) \) and \( M(.) \) respectively represent binary cross-entropy-with-logits loss and Multi-Assignment Sinkhorn-Knopp (discussed in Sec. 3.2), \( \alpha \) and \( \beta \) represents loss coefficients, and \( \text{sim}(a, b) = \exp(a^Tb/\tau||a||_2||b||_2) \), \( \tau \) is the temperature hyperparameter of the similarity metric.

Overall semantic structure preserving consistency loss for all the modalities is defined as:

\[
L_{sspc} = \sum_{m \in \{t,v,a\}} \sum_{n \in \{t,v,a\}} L(m, \hat{n}, \hat{z}_m, \hat{\tilde{z}}_m).
\]  

#### Contrastive Loss.

Following [11, 42], we also use contrastive loss to bring cross-modal embeddings of the same sample closer while pushing away embeddings from other sample. For this, we use 3 pairwise single-modality contrastive losses, \( L_{nce,tv}, L_{nce,ta}, L_{nce,va} \) between \((t, v), (t, a), (v, a)\) respectively. Specifically, we use Noise Contrastive Estimation [33] with temperature \( \kappa \) as shown in Eq. 4.

\[
L_{nce,xy} = -\log \frac{\exp(x^TY/\kappa)}{\sum_{i=1}^{N} \exp(x^TY^{(i)}/\kappa)}
\]

The overall contrastive loss for all modalities is defined as \( L_{nce} = \lambda_{tc} L_{nce,tv} + \lambda_{ta} L_{nce,ta} + \lambda_{va} L_{nce,va} \).

#### Overall Loss

The overall training objective is a combination of SSPC loss (Eq. 3) and contrastive loss (4):

\[
L_f = \lambda_{sspc} L_{sspc} + \lambda_{nce} L_{nce},
\]

where, \( \lambda_{sspc} \) and \( \lambda_{nce} \) are loss coefficients. By combining both losses, the model learns a more generic joint embedding space which preserves the modality-specific semantic structure by enforcing the anchor assignment similarity before and after feature projection and also brings representations of different modalities together by utilizing contrastive loss.

### 4. Experiments

#### 4.1. Experimental Setup

**Backbones.** For comparability, we follow the same setup of previous works [31, 39, 11, 42]. As visual backbone, we use a combination of 2D features from ResNet-152 [20] pretrained on ImageNet [14], and 3D features from ResNetXt101 [19] pretrained on Kinetics [10]. The text backbone is GoogleNews pretrained Word2vec model [32]. These backbones are fixed and not finetuned during training. Following [11, 42], we use a trainable CNN with residual layers as an audio backbone. We provide additional details in the Supplementary Sec. 5.

**Data Sampling.** We use a batch of 216 videos and randomly sample ten 8-second clips per video. If the sampled clip contains narration (95% clips), we use ASR time stamps to select clip borders. To disentangle high text-audio correlation in HT100M, we shift the audio clip randomly by 4 seconds with respect to the video and text boundaries.

**Projections.** Following [32, 11, 39], we use a gated linear projection [30] to project features into common token space, as well as to project resulting tokens into shared embedding space. We set the dimension of the common token space to 4096 and of the shared embedding space to 6144.
We use a single transformer block with hidden size of 4096 with 64 heads and an MLP size of 4096. We set the number of anchors $K$ as 64 and $K'$ as 32 with damping factor $\mu$ as 0.25. We train all models for 15 epochs using an Adam optimizer [24] with a learning rate of 5e-5, exponential decay of 0.9 and the temperature of cosine similarity ($\tau$) as 0.1. We maintain a memory-bank of size 5500 while performing Multi-SK. Following [1, 42], we set higher weight of loss terms that involve text-video in Eq. 3, i.e., for all text-video terms the weight is set to 1.0 and the rest of the weights are set to 0.1.

### 4.2. Datasets, Tasks, Metrics

**Pretraining Dataset.** We train our model on the HT100M dataset [31], which contains over 1 million instructional videos with automatically generated text narrations. The text narrations can be assumed to be noisy and to not always describe the video scene.

**Zero-shot Retrieval.** We use MSR-VTT [47] and YouCook2 [52] datasets to evaluate the zero-shot retrieval capability of our model. We report performance on 4 retrieval tasks: (i) Text-to-Video retrieval, (ii) Video-to-Text retrieval, (iii) Text-to-Video-Audio retrieval, (iv) Video-Audio-to-Text retrieval. The YouCook2 dataset contains cooking videos from YouTube with human-annotated clips ($\sim 2$ - 200 secs). For evaluation we use at maximum first 48 seconds of clip, since most clips are shorter than that. The MSR-VTT dataset contains human-annotated clips ($\sim 10$ - 30 secs) on various topics and provides captions with natural language sentences. Following [31, 39, 11, 42], to evaluate our model on MSR-VTT, we use the 1k set of test clips [49], and for YouCook2, we use 3,350 validation clips [31]. To perform ($t \rightarrow v + a$) retrieval, we compute similarities by dot product between a text query $t$ and all videos in the dataset using a averaged $v + a$ representa-

### 4.3. Comparison with State-of-the-art

**Zero-shot Retrieval Tasks.** In Tab. 1, we report the performance of the learned multi-modal representations on four zero-shot retrieval tasks, (i) Text-to-Video, (ii) Video-to-Text, (iii) Text-to-Video-Audio, (iv) Video-Audio-to-Text., on MSR-VTT and YouCook2 datasets. For a fair comparison, we only compare with models trained on all three modalities i.e. text, video and audio. In summary, our proposed method outperforms the current state-of-the-art methods by a noticeable margin on both the datasets. The results on the MSR-VTT dataset are particularly interesting since it demonstrates the generalizability of our model. In particular, HT100M consists of instructional videos and the textual descriptions are generated from audio using ASR. The text narration has noisy alignment with video and typically describes the steps in the video. YouCook2 shares

<table>
<thead>
<tr>
<th>Method</th>
<th>Retrieval Task</th>
<th>Dataset</th>
<th>Visual</th>
<th>Trainable BB</th>
<th>MSR-VTT</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t $\rightarrow$ v</td>
<td>HT100M</td>
<td>Res3D+ Faster R-CNN</td>
<td>BB</td>
<td>R@5↑</td>
<td>R@10↑</td>
</tr>
<tr>
<td>ActBERT [54]</td>
<td>t $\rightarrow$ v</td>
<td>HT100M</td>
<td>R512 + R(2+1)ID-34</td>
<td>✓</td>
<td>23.4</td>
<td>33.1</td>
</tr>
<tr>
<td>SupportSet [36]</td>
<td>t $\rightarrow$ v</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>23.0</td>
<td>31.1</td>
</tr>
<tr>
<td>HT100M [31]</td>
<td>t $\rightarrow$ v</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>21.2</td>
<td>29.6</td>
</tr>
<tr>
<td>AVLNet [39]</td>
<td>t $\rightarrow$ v</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>24.7</td>
<td>34.2</td>
</tr>
<tr>
<td>EAO [42]</td>
<td>t $\rightarrow$ v</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>24.6</td>
<td>35.3</td>
</tr>
<tr>
<td>Ours</td>
<td>t $\rightarrow$ v</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>26.4</td>
<td>35.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Retrieval Task</th>
<th>Dataset</th>
<th>Visual</th>
<th>Trainable BB</th>
<th>MSR-VTT</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVLNet [39]</td>
<td>v $\rightarrow$ t</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>27.2</td>
<td>35.7</td>
</tr>
<tr>
<td>EAO [42]</td>
<td>v $\rightarrow$ t</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>27.6</td>
<td>36.6</td>
</tr>
<tr>
<td>Ours</td>
<td>v $\rightarrow$ t</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>27.4</td>
<td>37.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Retrieval Task</th>
<th>Dataset</th>
<th>Visual</th>
<th>Trainable BB</th>
<th>MSR-VTT</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVLNet [39]</td>
<td>v $\rightarrow$ a</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>19.2</td>
<td>27.4</td>
</tr>
<tr>
<td>MCN [11]</td>
<td>v $\rightarrow$ a</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>25.2</td>
<td>33.8</td>
</tr>
<tr>
<td>EAO [42]</td>
<td>v $\rightarrow$ a</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>23.3</td>
<td>33.2</td>
</tr>
<tr>
<td>Ours</td>
<td>v $\rightarrow$ a</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>25.1</td>
<td>34.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Retrieval Task</th>
<th>Dataset</th>
<th>Visual</th>
<th>Trainable BB</th>
<th>MSR-VTT</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVLNet [39]</td>
<td>v $\rightarrow$ t</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>19</td>
<td>26.3</td>
</tr>
<tr>
<td>EAO [42]</td>
<td>v $\rightarrow$ t</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>21.8</td>
<td>31.4</td>
</tr>
<tr>
<td>Ours</td>
<td>v $\rightarrow$ t</td>
<td>HT100M</td>
<td>R152 + RX101</td>
<td>✓</td>
<td>24</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 1: Zero-shot Retrieval results on MSR-VTT/YouCook2. For fair comparison, we compare with models trained on text, video and audio. Retrieval column represents the evaluation task. BB=Backbone. **Bold**, **underline** represent highest and second-highest scores.
domain similarity with HT100M as its videos are instructional format, and the text descriptions corresponds to specific steps in the recipe. In contrast, MSR-VTT is not restricted to instructional videos and the text is typically a single sentence caption describing the whole video. As a result, there is a distributional shift between HT100M and MSR-VTT. Therefore, zero-shot retrieval on MSR-VTT has to overcome distributional shift, which is a crucial requirement for practical deployment. Our proposed method shows relatively higher improvement on this task validating the effectiveness of anchor-based learning.

In case of text-to-video-audio retrieval, our method improves the Median and Mean rank of the baseline [42] by 3% on MSR-VTT along with gains on recall metrics R@5 and R@10. For video-audio-to-text retrieval, our methods performs very well on MSR-VTT with 3% improvement on MeanR, 2.2% improvement on R@5. We also observe similar improvements on the YouCook2 dataset. Similarly, our proposed method outperforms the current state-of-the-art on most of the metrics for text-video-retrieval video-text-retrieval. Additionally, to compare our approach with text-video only model, we train our model on video and text to compare the performance with state-of-the-art in Sec 1 of the Supplementary.

Zero-shot Full Video Retrieval. In Tab. 3, we report results for text-to-full-video retrieval task. Our approach outperforms prior works by 2.2%, 2.7% on R@1 an R@5 respectively.

Retrieval after Fine-tuning. We also evaluate the retrieval performance of our model after fine-tuning on the downstream datasets as shown in Tab. 2. For a fair comparison, we only report the baselines that use the same training split. We outperform state-of-the-art consistently on all the metrics and on both MSR-VTT & YouCook2 datasets as shown in Tab. 2.

4.4. Ablation Studies

First we analyze the impact of the proposed components and report results in Tab. 4, followed be effect of number of anchors (K) & (K') in Tab. 5.

Table 3: Zero-Shot Text-to-Full Video retrieval on YouCook2. Averaging across captions is used to obtain video-level predictions.

Table 4: Ablation studies showing the impact of various components for zero-shot retrieval task. Recon. = Reconstruction Loss, CM SSPC = Cross-Modal SSPC Loss, SK = Sinkhorn-Knopp, CL=Contrastive Loss.

Effect of Proposed Components. We report the results for this ablation in Tab. 4. In first row, we report results using reconstruction loss and contrastive loss, we notice that using reconstruction loss reduces the performance by 2% on all metrics indicating the effectiveness of the proposed SSPC loss. In the second row, we report results without our proposed cross-modal SSPC loss (‘w/o CM SSPC’). We employ this loss to bring the cross-modal representations closer in the joint embedding space to obtain better performance in zero-shot cross-modal tasks. We notice that removing the cross-modal SSPC loss drastically decreases the zero-shot retrieval performance on both MSR-VTT and YouCook2 datasets, 2.4% and 2.5% drop in R@5 performance respectively. This empirically validates the effectiveness of the cross-modal SSPC loss in obtaining better cross-modal representations.

In the third row, we analyze the effect of the proposed SSPC loss (‘w/o SSPC’). We apply the SSPC loss to retain modality-specific semantic structure from the pre-trained models in the joint embedding space. To investigate its impact, we remove the anchor consistency between the modality-specific and joint embedding spaces. Instead, we enforce the anchor assignments before Multi-SK to be consistent with the Multi-SK optimized anchor as-

<table>
<thead>
<tr>
<th>Method</th>
<th>Aggregation</th>
<th>MSR-VTT</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HT100M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIL-NCE [29]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCN [11]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAO [42]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>MSR-VTT</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recon. + CL</td>
<td>23.1</td>
<td>32.4</td>
</tr>
<tr>
<td>w/o CM SSPCL</td>
<td>22.7 31.8</td>
<td></td>
</tr>
<tr>
<td>w/o SSPC</td>
<td>23.8</td>
<td>33.3</td>
</tr>
<tr>
<td>Modified SK</td>
<td>23.4</td>
<td>31.3</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>25.1</strong></td>
<td><strong>34.5</strong></td>
</tr>
</tbody>
</table>

Table 4: Ablation studies showing the impact of various components for zero-shot retrieval task.

Loss, CM SSPC = Cross-Modal SSPC Loss, SK = Sinkhorn-Knopp, CL=Contrastive Loss.
Figure 2: Anchor assignments illustrated in this figure demonstrate the visual similarity between video samples from two related categories i.e. (a) Holidays and Tradition and (b) Food and Entertaining; and two different categories i.e. (d) Personal Care ans Style and (e) Finance and Business within the HowTo100M dataset. The visual similarity is also reflected in video anchor assignments (c) as most assigned anchors are similar with minor differences in assignments, thereby showcasing the flexibility and effectiveness of our approach. (d), (e) samples look very different and therefore the anchor assignments are also very different as shown in (f). Green cell → Anchor assigned, Yellow → Anchor not assigned. Difference in anchor assignments indicated in red. 

<table>
<thead>
<tr>
<th>$K$</th>
<th>$K'$</th>
<th>MSR-VTT</th>
<th>YouCook2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@5↑</td>
<td>R@10↑</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>23.1</td>
<td>32</td>
</tr>
<tr>
<td>32</td>
<td>16</td>
<td>23.2</td>
<td>32.1</td>
</tr>
<tr>
<td>64</td>
<td>16</td>
<td>23.3</td>
<td>31.8</td>
</tr>
<tr>
<td>64</td>
<td>48</td>
<td>23.7</td>
<td>32.1</td>
</tr>
<tr>
<td>64 (Ours)</td>
<td>32</td>
<td>25.1</td>
<td>34.5</td>
</tr>
</tbody>
</table>

Table 5: Effect of different # of anchors on zero-shot retrieval. $K$ → # of anchors and $K'$ → # of selected anchors, respectively.

assignments from the same embedding space. We also apply the cross-modal SSPC loss to only isolate the impact of the proposed SSPC loss across feature projections. We observe that removing the anchor consistency between the modality-specific and joint embedding spaces reduces the zero-shot retrieval performance by a significant margin ($\sim 1.5\%$ drop in R@5 performance), indicating the importance of SSPC loss in maintaining modality-specific semantic structure in the joint embedding space for better performance.

Finally, in the fourth row of Tab. 4, we analyse the effectiveness of the proposed Multi-Assignment Sinkhorn-Knopp algorithm. To this end, we modify the vanilla Sinkhorn-Knopp [13] to obtain multiple anchor assignments per sample (refer to Sec. 5 of the Supplementary for details). We notice that the modified Sinkhorn-Knopp does not perform well with a drop in performance ($\sim 1.6\%$ drop in R@5) on both MSR-VTT and YouCook2 datasets on all the metrics indicating the significance of our proposed MULTI-SK algorithm.

**Effect of Number of Anchors.** To analyze the effect of the number of anchors, we conduct experiments with different number of anchors, $K$, and different number of selected anchors, $K'$. We report the results in Tab. 5. We conduct experiments on both MSR-VTT and YouCook2 datasets. From the top half of Tab. 5, we observe that the performance of our proposed method improves with the increasing number of anchors. This is to be expected since a higher number of anchors have a higher representation learning capacity. We also observe that the method performs reasonably well even with a very small number of anchors showing the general effectiveness of our proposed solution. We notice that the performance of the proposed method improves as we select more anchors as shown in bottom half of Tab. 5. However, selecting a very large number of anchors (48 out of 64 for this experiment) adds more constraints leading to poor performance.

### 4.5. Qualitative Analysis

Here, first we present a fine-grained visual analysis of the learned anchors, followed by qualitative retrieval results.
We show fine-grained analysis of the learned anchors in Fig. 2. For the purpose of this analysis, we visualize the anchor assignments as binary assignments. However, during training we use soft anchor assignments. In Fig. 2(c), we compare the anchor assignments for samples from similar categories. It can be seen that the videos are visually similar even though they belong to different categories and the anchor assignments for these two examples are able to capture the sample similarity. In Fig. 2(f), we compare the anchor assignments for videos from different categories and it can be seen that the anchor assignments are very different as expected. This further validates our claim that our proposed method can assign semantically meaningful anchors without any explicit supervision. Further, we show qualitative retrieval comparison with EAO [42] in Fig. 3. We present more qualitative analysis in Supplementary Sec. 4.

5. Conclusion

We proposed a novel approach that preserves the modality-specific semantic relationship between samples in the joint multi-modal embedding space. To this end, we propose a flexible sample relationship modeling approach by assigning multiple anchors to each sample, which captures both shared and unique aspects of samples. To obtain these assignments, we develop a novel Multi-Assignment Sinkhorn-Knopp (Multi-SK) algorithm, and also utilize the proposed anchor consistency loss to learn these anchors. Our qualitative results demonstrate that our learnt anchors correspond to meaningful semantic concepts. Our extensive experimentation demonstrates that the proposed approach improves generalizability by outperforming state-of-the-art methods on both in- and out-of-domain datasets. We also show that our method achieves state-of-the-art performance on multiple zero-shot tasks, and also outperforms when fine-tuned on downstream datasets.

Acknowledgements

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-19-1-0356. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Nina Shvetsova is supported by German Federal Ministry of Education and Research (BMBF) project STCL - 01IS22067.

References


[27] Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2


[31] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching


