Learning a Task-Adaptive Low-Dimensional Semantic Space for Improved Visual Classification

Nikalal Kaldera Wanqing Li Chris Bunn Philip O. Ogunbona
School of Computing and Information Technology
University of Wollongong
Wollongong, NSW 2500, Australia

Abstract

Semantic visual classification commonly aligns visual embeddings with semantic spaces induced by pretrained text encoders. However, such spaces are typically high-dimensional, weakly aligned with visual feature geometry, and insufficiently discriminative for downstream classification tasks. We propose TaSS, a framework for learning a task-adaptive low-dimensional semantic space that preserves semantic structure while explicitly optimizing class separability. TaSS is constructed through a parameterized geometry transformation of pretrained text embeddings that enlarges inter-class margins via distance reshaping and improves intra-class compactness through prototype-preserving semantic constraints. The resulting semantic space maintains meaningful semantic relationships while producing a discriminative geometry tailored for classification. Frozen visual encoder features are subsequently aligned to TaSS using a lightweight MLP, enabling improved discriminative learning without fine-tuning pretrained backbones. We further provide theoretical analysis showing that TaSS preserves semantic consistency while enforcing enhanced inter-class separation and controlled intra-class variance. Extensive experiments on skeleton-based human action recognition, video classification, and image classification benchmarks demonstrate consistent improvements over existing semantic alignment approaches, with gains exceeding 3 percentage points on average and up to 8.5 points on challenging datasets.

1 Introduction

Semantic visual classification commonly relies on aligning visual representations with semantic embeddings derived from pretrained language models such as CLIP Radford et al. [2021]. In this paradigm, semantic embeddings encode relationships between classes independently of visual observations and provide transferable supervision for downstream recognition tasks. Existing approaches typically map visual features into a semantic space induced directly by pretrained text encoders (Fig. 1a).

Despite their strong generalization ability, pretrained semantic spaces are not necessarily optimal for discriminative visual classification. First, they are often high-dimensional and geometrically mismatched with compact visual representations extracted by modern vision encoders. Second, semantic proximity in language space does not always correspond to discriminative separability for visual recognition, causing semantically related classes to overlap. Third, directly inheriting the geometry of pretrained language embeddings provides limited control over intra-class structure and inter-class margins in downstream tasks Schall et al. [2025]. Consequently, semantic alignment may preserve linguistic relationships while failing to produce sufficiently discriminative representations for classification.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

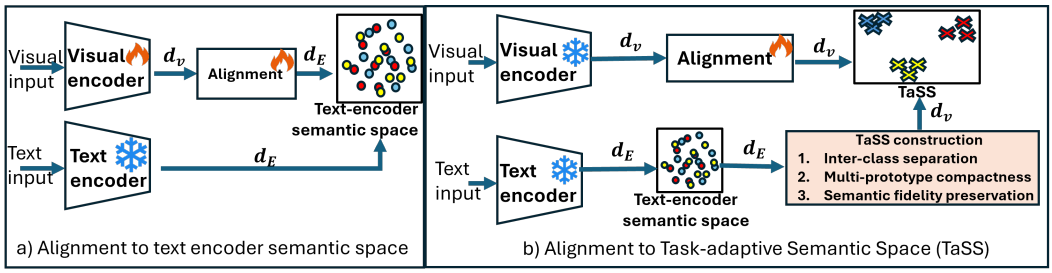


Figure 1: Overview of text-guided visual classification. (a) Alignment to the original text-encoder semantic space that is dimensionally mismatched with visual features is not optimized for discrimination, limiting class separability. (b) Our method aligns frozen visual features to TaSS constructed from the text-encoder space, with controllable dimensionality, enhanced inter-class separation, multi-prototype compactness, and semantic preservation.

Table 1: Comparison of dimensionality reduction methods.

Method	Dim. Control	Separation	Compactness	Semantic Fidelity	Supervised
MDS	High	Low	Low	Yes	No
t-SNE	High	Low	Medium	No	No
UMAP	High	Medium	Medium	No	No
LDA	Low	High	High	No	Yes
TaSS (Ours)	High	High	High	Yes	Yes

This motivates the following question: ***Can we learn a semantic space that simultaneously preserves semantic relationships and optimizes discriminative geometry for visual classification?*** More specifically, we seek a semantic space satisfying three properties: (a) *dimensional compatibility with visual encoder features*, (b) *enhanced inter-class separation and intra-class compactness*, and (c) *preservation of the semantic structure inherited from pretrained text embeddings*.

Conventional dimensionality reduction methods only partially address these objectives. While they may reduce dimensional mismatch, they generally lack mechanisms for jointly preserving semantic consistency and enforcing discriminative class geometry, as summarized in Table 1.

To address this limitation, we propose TaSS, a framework for learning a discriminative low-dimensional semantic space tailored for visual classification. TaSS is constructed through a parameterized transformation of pretrained semantic embeddings that explicitly reshapes semantic geometry while preserving semantic fidelity. In particular, inter-class discrimination is enhanced through a concave distance transformation that enlarges pairwise margins between class-level embeddings. To improve intra-class compactness without collapsing meaningful semantic variation, we introduce a prototype-preserving formulation that identifies multiple semantic modes within each class and constrains their scatter around class centroids in the learned space. Together, these mechanisms produce a semantic representation that preserves meaningful class relationships while improving discriminative structure.

Once the TaSS is constructed, frozen visual features extracted from image, video, or motion encoders are aligned to the learned semantic space through a lightweight MLP without fine-tuning the underlying backbone networks (Fig. 1b). This decouples semantic-space learning from visual feature extraction and enables improved classification performance while retaining the generalization capability of pretrained encoders.

We further provide theoretical analysis showing that the proposed framework preserves semantic consistency while enforcing enhanced inter-class separation and bounded intra-class variance in the learned semantic space.

Extensive experiments on skeleton-based human action recognition, video classification, and image classification benchmarks demonstrate the effectiveness and generality of TaSS across multiple modalities and architectures. The proposed framework consistently improves classification performance over existing semantic alignment methods, achieving gains exceeding 3 percentage points on average and up to 8.5 points on challenging benchmarks.

The main contributions of this work are summarized as follows:

- We propose TaSS, a framework for learning a discriminative low-dimensional semantic space for visual classification that jointly enforces dimensional compatibility, semantic consistency, inter-class separation, and intra-class compactness.
- We introduce a geometry-transformation and prototype-preserving formulation that reshapes pretrained semantic embeddings into a task-adaptive discriminative semantic space.
- We provide theoretical analysis demonstrating that TaSS preserves semantic structure while improving class separability and controlling intra-class variance.

2 Related work

Dimensionality Reduction: Existing methods often learn discriminative low-dimensional representations by increasing class separability and reducing class variance, but they do not preserve semantics, as shown in Table 1. For instance, Linear Discriminant Analysis (LDA) Fisher [1936] does not preserve global geometry, which is important for semantic preservation. LFDA Sugiyama [2007], MFA Yan et al. [2006], and Graph-Regularized LDA Cai et al. [2007] prioritize local discrimination and often lose global structure. Supervised Isomap Geng et al. [2005] enhances between-class separation but provides limited compactness and geometry preservation. Conversely, MDS Kruskal [1964] preserves global pairwise geometry but does not improve separation or compactness, while UMAP McInnes et al. [2018] and t-SNE Van der Maaten and Hinton [2008] *neither guarantee global geometry nor explicitly enhance class discrimination*.

Metric and prototype-based learning: Metric and prototype-based learning improves classification through compact intra-class and separated inter-class embeddings Kulis [2013]. Supervised contrastive learning Khosla et al. [2020] pulls same-class samples together and pushes negative samples apart, while Multi-Similarity Wang et al. [2019] and Circle Loss Sun et al. [2020] focus on hard pairs. Proxy Anchor Loss Kim et al. [2020] learns class proxies, and ArcFace Deng et al. [2019] adds an angular margin for sharper boundaries. However, *these methods improve discrimination without preserving semantics*.

Text-guided visual classification. Text-guided classification aligns visual and textual features for semantic motion, video, and image recognition. In image classification, CLIP Radford et al. [2021] aligns image-text features via contrastive learning, while CoOp Zhou et al. [2022b], CoCoOp Zhou et al. [2022a], MaPLe Khattak et al. [2023a], and PromptSRC Khattak et al. [2023b] adapt and regularize textual semantics for downstream tasks. Tip-Adapter Zhang et al. [2022] and WATT Osowiechi et al. [2024] further improve classification using multiple textual descriptions. For motion classification, ActionCLIP Wang et al. [2021] extends CLIP to video recognition, while GAP Xiang et al. [2023] and CrossGLG Yan et al. [2024] align skeleton features with textual semantics. In zero-shot recognition, PURLS Zhu et al. [2024], Neuron Chen et al. [2024b], and FS-VAE Wu et al. [2025] leverage motion-phase or frequency-aware text alignment to distinguish similar actions. However, *these methods mainly improve semantic alignment without explicitly enhancing target-space discrimination*, which limits class separability.

3 Problem formulation

A semantic space is a high dimensional vector space in which linguistic (or other meaningful) items are represented as vectors, such that semantic relatedness is captured by geometric proximity (e.g. distance or angle) between those vectors Ethayarajh [2019], Karlgren and Kanerva [2021]. Let $\mathcal{C} = \{1, \ldots, C\}$ denote a set of semantic classes. A collection of semantic embeddings is denoted as $\{(e_i, c_i)\}_{i=1}^{n}$, where $c_i \in \mathcal{C}$ is its associated class, and $E = \{e_i\}_{i=1}^{n}$, $e_i \in \mathbb{R}^{d_E}$, denote the set of embeddings.

Let $\mathcal{E} \subseteq \mathbb{R}^{d_E}$ denote the original semantic space produced by a pretrained text encoder, where d_E is the semantic embedding dimension. Let $\rho_E : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_{\geq 0}$ denote the semantic dissimilarity metric induced by the original semantic embeddings. The objective is to learn a mapping, $g_\theta : \mathbb{R}^{d_E} \to \mathbb{R}^{d_v}$ that maps the original semantic embeddings into the TaSS. Here, d_v denotes both the visual feature dimension and the TaSS dimension, since the TaSS is constructed to serve as the direct alignment target for visual features. The TaSS is designed to satisfy three properties: dimensional compatibility

with visual features, enhanced inter-class separation and intra-class compactness, and preservation of semantic fidelity.

Definition 3.1 (Semantic Space with Semantic Fidelity)**.** Let $\mathcal{P} = \{(c, c') : 1 \leq c < c' \leq C\}$ denote the set of all unordered class pairs. A target space $(\mathcal{C}, \rho_M)$ is said to maintain *semantic fidelity* with respect to $(\mathcal{C}, \rho_E)$ if, for any $(c, c'), (k, k') \in \mathcal{P}$,

$$\rho_E(c, c') < \rho_E(k, k') \implies \rho_M(c, c') < \rho_M(k, k'). \tag{1}$$

This definition yields a task-agnostic criterion: the relative order of class dissimilarities induced by ρ_E is preserved under ρ_M, even when the absolute geometry of the space changes.

Condition 1 (Sufficient Condition for Semantic Fidelity)**.** *Let $t_{cc'}$ denote the desired target inter-class distance between classes c and c'. The condition*

$$\Delta_{\min} = \min_{\substack{(c,c'),(k,k') \in \mathcal{P} \\ t_{cc'} \neq t_{kk'}}} \left| t_{cc'} - t_{kk'} \right|, \tag{2}$$

denote the minimum separation between distinct classes in the target space. The worst case approximation error with respect to the target distances is defined as,

$$\varepsilon_\infty = \max_{(c,c') \in \mathcal{P}} \left| \rho_M(c, c') - t_{cc'} \right|. \tag{3}$$

If $\varepsilon_\infty < \frac{\Delta_{\min}}{2}$, then the derived semantic space $(\mathcal{C}, \rho_M)$ maintains semantic fidelity with respect to $(\mathcal{C}, \rho_E)$.

4 Learning a task-adaptive semantic space (TaSS)

Semantic geometry enhancement: To improve inter-class separation while preserving semantic fidelity, a target inter-class distance configuration is constructed to expand distances between class level representations. The idea is to apply an order preserving transformation to the distances of the semantic space, thereby increasing separation among nearby classes without altering their relative semantic ordering.

This can be achieved by applying an increasing concave transformation to the reference inter-class distances $\delta_{cc'} = \rho_E(c, c')$ to obtain the target distances $t_{cc'}$:

$$t_{cc'} = f_a(\delta_{cc'}) = \delta_{\max} \left(\frac{\delta_{cc'}}{\delta_{\max}} \right)^a, \tag{4}$$

where $a \in (0, 1]$ controls the degree of concavity and $\delta_{\max} = \max_{c<c'} \delta_{cc'}$. Since f_a is strictly increasing, the relative ordering of inter-class distances is preserved, thereby maintaining semantic fidelity.

In addition, for $a < 1$, the ratio $(x/\delta_{\max})^{a-1}$ decreases with x, causing smaller distances to be amplified more than larger ones. This selectively increases separation among semantically similar classes while minimally affecting dissimilar pairs.

Learning the target semantic geometry: Let $m_i = g_\theta(e_i) \in \mathbb{R}^{d_v}$ denote the mapped semantic embedding in the TaSS. Class level representations are obtained by computing centroids over the embeddings belonging to each class, denoted by μ_c^M. These class centroids lie on the unit hypersphere $\mathbb{S}^{d_v - 1}$ and induce the learned TaSS metric ρ_M via angular distance: $\rho_M(c, c') = \arccos(\mu_c^M \cdot \mu_{c'}^M)$

To align the learned class centroids with the target inter-class distance configuration $\{t_{cc'}\}$, a stress based objective is minimized:

$$\mathcal{L}_{\text{sem_geo}} = \frac{1}{|\mathcal{P}|} \sum_{(c,c') \in \mathcal{P}} \left(\arccos(\mu_c^M \cdot \mu_{c'}^M) - t_{cc'} \right)^2. \tag{5}$$

This objective penalizes deviations between the learned and target inter-class distances, driving the learned semantic metric to approximate the target distance configuration.

Compactness enhancement via multi-prototype regularization: While the separation objective enhances inter-class separation, it does not constrain within-class scatter, which can lead to overlap between nearby classes. To address this issue, we introduce a compactness regularization that controls intra-class geometry while preserving meaningful variation or style within each class.

We represent each class using K prototypes. Each prototype captures one mode of semantic variation, such as styles, within the class. The prototypes are obtained through clustering, such as k-means, to the original semantic embeddings of each class in $\mathbb{R}^{d_E}$, and are then mapped into the TaSS.

To enforce compactness, two constraints are introduced. First, each embedding $m_i = g_\theta(e_i)$ is forced to remain close to its assigned prototype π_i as,

$$\mathcal{L}_{\text{local}} = \frac{1}{n} \sum_{i=1}^{n} \left\| m_i - p^M_{c_i,\pi_i} \right\|_2^2, \tag{6}$$

where c_i is the class label of embedding e_i, and $p^M_{c_i,\pi_i}$ denotes the TaSS prototype assigned to m_i.

Second, the prototype centers of each class are regularized to remain compact around the class centroid μ^M_c. Specifically, we constrain the centroid-to-prototype distances to remain smaller than the global classwise separation in the target space, thereby reducing overlap between classes. This is achieved by assigning each prototype a target centroid-to-prototype distance $r_{c,j}$, determined in the original semantic space. The maximum radius is scaled to half the minimum target inter-class centroid separation, thereby limiting intra-class prototype spread relative to inter-class separation and reducing potential cross-class overlap:

$$\mathcal{L}_{\text{proto}} = \frac{1}{C} \sum_{c \in \mathcal{C}} \frac{1}{K} \sum_{j=1}^{K} \left(\left\| p^M_{c,j} - \mu^M_c \right\|_2 - r_{c,j} \right)^2. \tag{7}$$

Together, a semantic variation or style is constructed, $\mathcal{L}_{\text{sem_sty}} = \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{proto}}$

Total Objective & Algorithm: The TaSS mapping g_θ is trained with $\mathcal{L}_{\text{TaSS}} = \mathcal{L}_{\text{sem_geo}} + \lambda \mathcal{L}_{\text{sem_sty}}$, where $\lambda = 1$ balances semantic geometry and style preservation. Algorithmic details are given in Appendix section B.

5 Theoretical properties of the TaSS

Proposition 1 (Semantic fidelity) When Condition 1 holds, the derived semantic space $(\mathcal{C}, \rho_M)$ reflects the relative ordering of inter-class semantic dissimilarities induced by the original semantic metric ρ_E.
Proposition 2 (Selective separation enhancement) For concavity parameter $a < 1$, the target distance transformation expands smaller reference inter-class distances proportionally more than larger ones, thereby increasing separation among semantically similar classes.
Proposition 3 (Effect of compactness regularization) Under the assumption that multiple semantic embeddings associated with the same class capture stylistic variations, prototype-based compactness regularization acts on intra-class dispersion and introduces no explicit cross-class interactions.
Formal statements and proofs are provided in Appendix section A.

6 Application to visual classification

To construct a TaSS for a visual task, l textual descriptions are generated per class using a large language model to capture class specific variations. The descriptions are encoded using a pretrained text encoder such as CLIP Radford et al. [2021], BERT Devlin et al. [2019], BGE Xiao et al. [2024], and the resulting embeddings are processed using TaSS Algorithm (1). Prompting and text encoding details are provided in Appendix section D.

Let f_ϕ denote a frozen visual encoder that maps input visual content x to a feature vector $v = f_\phi(x) \in \mathbb{R}^{d_v}$. The visual-to-TaSS alignment function is denoted by $h_\psi : \mathbb{R}^{d_v} \to \mathbb{R}^{d_v}$, which aligns visual features to K prototypes of the TaSS via h_ψ, which is an MLP. The TaSS is constructed with the same dimensionality as the visual encoder output so that visual features and TaSS prototypes lie in the same space $\mathbb{R}^{d_v}$.

Multi-prototype alignment: Instance-level visual features generated by f_ϕ are aligned with K prototypes of each class in the TaSS using a learnable mapping function. Let $P_c = \{p_{c,1}, \ldots, p_{c,K}\} \subset \mathbb{S}^{d_v-1}$ denote the prototypes associated with class c in the TaSS. For notational simplicity, we omit the superscript M in this section and denote TaSS prototypes by $p_{c,j}$. Given a visual feature v, the alignment function maps it to the TaSS as $z = h_\psi(v),\ z \in \mathbb{S}^{d_v-1}$.

For a visual embedding z with ground-truth class c, the positive prototype p^+ is selected as the most similar prototype from the corresponding class, i.e. $p^+(z, c) = \arg\max_{p \in P_c}\ z^\top p$.

All prototypes belonging to the remaining classes are considered negatives. Based on this, the probability of assigning z to its positive prototype is defined using an InfoNCE-style objective:

$$p\big(p^+ \mid z\big) = \frac{\exp\big(z^\top p^+/\tau\big)}{\exp(z^\top p^+/\tau) + \sum_{c' \neq c} \sum_{j=1}^{K} \exp(z^\top p_{c',j}/\tau)}, \tag{8}$$

where $\tau > 0$ is a temperature parameter. The visual-semantic alignment loss is then given by $\mathcal{L}_{\text{vis}} = -\frac{1}{N} \sum_{i=1}^{N} \log p\big(p_i^+ \mid z_i\big)$, where N is the number of samples.

Inference: At inference, classification is performed by nearest-prototype matching based on maximum semantic similarity: $\hat{y} = \arg\max_{c \in \mathcal{C}} \max_{1 \leq j \leq K}\ z^\top p_{c,j}$, where $z = h_\psi(f_\phi(x))$ is the visual embedding aligned to the TaSS, and $p_{c,j}$ denotes the j-th TaSS prototype of class c.

7 Experiments

The proposed method has been validated on a variety of classification datasets. They are **Skeleton-based human action recognition** (HAR) on *NTU RGB+D* Shahroudy et al. [2016] (NTU-60), a 3D action dataset with 60 actions, performed with 40 different subjects; *NTU RGB+D 120* Liu et al. [2020] (NTU-120), an extension of the NTU-60 with additional 60 actions, with 106 different subjects; **Video classification** on *HMDB51* Kuehne et al. [2011] contains 6,766 video clips of 51 action categories, and on *Aridv1.5* Xu et al. [2021] contains 3,784 video clips of 11 action categories; **Image classification** on *CIFAR-10* Krizhevsky [2009], a widely used image classification dataset with 60,000 low-resolution RGB images across 10 classes, *CUB200* Wah et al. [2025], a widely used fine-grained image classification dataset with 200 bird species, and *Oxford-IIIT Pet* Parkhi et al. [2012] is a fine-grained image classification dataset of 37 pet categories including dog and cat breeds.

Implementation: All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU. Class-level text descriptions are generated using GPT such as GPT-5 Singh et al. [2025], Gemini Anil et al. [2023], Claude Anthropic [2024], and encoded with a pretrained text encoders such as CLIP Radford et al. [2021], BERT Devlin et al. [2019] and BGE Xiao et al. [2024]. The prompt template used for text generation is provided in Appendix section D. For evaluation, pretrained weights from the original repositories were used for all encoders. Specifically, USDRL Wang et al. [2025] and MAMP Mao et al. [2023] were used as motion encoders; VideoMAEv2 Wang et al. [2023] and SMILE Thoker et al. [2025] with ViT-B/16 were used as video encoders; and DINOv3 Siméoni et al. [2025] with ViT-S/16 and MAE He et al. [2022] with ViT-B/16 were used as image encoders.

Results: The effectiveness of TaSS alignment is evaluated across tasks, datasets, and encoder families, as summarized in table 2. Four settings are compared: *Linear probe*, a supervised linear classifier on frozen features without semantic alignment; *TaSS linear*, a linear projection to TaSS; *Text-Enc. MLP Align.*, MLP alignment to the original text-encoder semantic space; and *TaSS MLP Align.*, MLP alignment to TaSS.

***Comparison with the original text-encoder semantic space*:** To determine whether the gains come from semantic supervision alone or from constructing a discriminative target semantic space, we align frozen visual features using the same projection layer to either the original text-encoder semantic space (*Text-Enc. MLP Align.*) or TaSS (*TaSS MLP Align.*). *TaSS MLP Align.* consistently outperforms *Text-Enc. MLP Align.* across all datasets and encoders; for example, it improves NTU120 accuracy with the frozen USDRL encoder by 1.3 percentage points. Notable gains occur for semantically similar action pairs, such as "put object into bag"/"take object out of bag" and "thumbs up"/"thumbs

Table 2: TaSS alignment improves visual classification across HAR, video, and image classification tasks, datasets, encoder architectures. Each row evaluates a frozen encoder under four settings: linear probing, linear projection to TaSS, MLP alignment to the original text-encoder semantic space, and MLP alignment to TaSS. The final column reports the TaSS gain over MLP alignment to the original text-encoder semantic space on the hardest 30% of classes. Here SSL, SL, Trans., GCN, and Align. denote self-supervised learning, supervised learning, Transformer, graph convolutional network, and alignment, respectively. For Text-Enc. MLP Align. and TaSS MLP Align., results are reported as mean $\pm$ standard deviation over three independent seeds.

Task	Dataset	Encoder	Publication	Arch.	Pretrain	Top-1 Accuracy				Hard-class Gain
						Linear Probe	TaSS Linear	Text-Enc. MLP Align.	TaSS MLP Align.	
HAR	NTU60	USDRL	PAMI'25	Trans.	SSL	84.2	85.2	84.9±0.1	**85.7±0.2**	2.6
		MAMP	ICCV'23	Trans.	SSL	84.9	87.1	85.2±0.2	**88.2±0.2**	6.3
		CTR-GCN	ICCV'21	GCN	SL	89.9	90.6	89.9±0.1	**90.9±0.2**	1.1
	NTU120	USDRL	PAMI'25	Trans.	SSL	74.9	76.5	75.4±0.1	**76.7±0.3**	4.0
		MAMP	ICCV'23	Trans.	SSL	78.6	80.8	79.1±0.3	**81.1±0.1**	4.1
		CTR-GCN	ICCV'21	GCN	SL	84.9	85.7	85.3±0.1	**85.9±0.2**	1.4
Video	HMDB51	SMILE	CVPR'25	Trans.	SSL	70.8	71.5	71.4±0.1	**72.0±0.2**	0.9
		VideoMAEv2	CVPR'23	Trans.	SSL	83.2	84.8	84.0±0.1	**84.9±0.1**	0.9
	Aridv1.5	SMILE	CVPR'25	Trans.	SSL	71.5	76.7	74.2±0.3	**77.2±0.4**	1.5
		VideoMAEv2	CVPR'23	Trans.	SSL	79.0	79.9	79.2±0.2	**80.1±0.1**	0.8
Image	CIFAR10	DINOv3	arXiv'25	Trans.	SSL	96.6	97.7	97.4±0.2	**98.1±0.3**	0.7
		MAE	CVPR'22	Trans.	SSL	84.2	86.0	84.8±0.1	**89.2±0.2**	6.5
	CUB-200	DINOv3	arXiv'25	Trans.	SSL	84.8	88.2	85.4±0.2	**88.2±0.3**	8.5
		MAE	CVPR'22	Trans.	SSL	30.7	38.5	35.5±0.2	**40.1±0.1**	8.3
	Oxford -IIIT Pet	DINOv3	arXiv'25	Trans.	SSL	94.1	95.0	94.5±0.1	**95.0±0.2**	1.6
		MAE	CVPR'22	Trans.	SSL	80.9	82.8	82.9±0.3	**84.4±0.4**	1.3

down", which are weakly separated in the original text-encoder space. By enhancing inter-class separation while preserving semantic fidelity, TaSS provides a more discriminative alignment target.

For fine-grained image classification, TaSS improves over *Text-Enc. MLP Align.* by 2.8 percentage points on CUB200 with the frozen DINOv3 encoder. Large gains appear among visually and semantically similar bird groups, such as gulls (e.g., "California gull", "Herring gull"), terns (e.g., "Arctic tern", "Black tern"), and sparrows (e.g., "Baird sparrow", "Brewer sparrow"). These species share common visual and semantic attributes and are therefore weakly separated in the original text-encoder semantic space; such fine-grained distinctions particularly benefit from the proposed discriminative alignment.

***Hard-class performance*:** Hard-class gains are reported for the hardest 30% of classes, defined as those with the lowest per-class accuracy under baseline *Text-Enc. MLP Align.*. As shown in table 2, *TaSS MLP Align.* improves over *Text-Enc. MLP Align.* on these hard classes, with maximum gains of 6.3, 1.5, and 8.5 percentage points on HAR, video, and image classification, respectively. These gains demonstrate that TaSS provides a more discriminative semantic space for ambiguous classes.

***Comparison with metric-learning baselines*:** Since TaSS aims to improve discriminative visual representations, it is compared with SupCon Khosla et al. [2020], Multi-Similarity Wang et al. [2019], Circle loss Sun et al. [2020], and ArcFace Deng et al. [2019] under the same frozen-DINOv3 setting on CUB200. This comparison tests whether learning a discriminative embedding space from frozen visual features alone can match TaSS alignment. Unlike TaSS, these baselines use class-label supervision to improve discrimination without explicitly retaining class-wise semantics.

As shown in table 3, all metric-learning baselines improve over the linear probe, confirming the value of discriminative supervision. However, TaSS achieves the best Top-1 accuracy, outperforming the strongest metric-learning baseline, Circle loss, by 1.5 percentage points. This suggests that the gain is not only from enforcing compactness and separation using visual labels, but also from aligning visual features to a semantic target space designed to improve discrimination.

Table 3: Comparison with metric-learning methods on CUB-200 using DINOv3.

Method	Top-1 accuracy
Linear probe	84.8
SupCon	86.3
Multi-Similarity	86.4
Circle	86.7
ArcFace	86.2
TaSS (ours)	**88.2**

Table 4: Projection design ablation. h, LN and d_v denote hidden dimension, layer normalization and visual feature dimension.

Variant	Design	h	LN	HAR	Image cls.
Baseline	–	–	–	84.9	84.2
Linear	Linear	–	No	87.1	86.0
MLP	MLP-2	d_v	No	87.5	87.2
	LN-MLP-2	d_v	Yes	87.9	88.3
	LN-MLP-2	$d_v/2$	Yes	87.8	87.0
	LN-MLP-3	d_v	Yes	**88.2**	**89.2**
	LN-MLP-4	d_v	Yes	88.0	89.1

MLP-n denotes n linear layers with GELU between consecutive linear layers.

Table 5: Effect of the number of prototypes.

Prototype count	Top-1	
	HAR	Image cls.
2	87.6	88.5
3	87.8	**89.2**
4	**88.2**	**89.2**
5	88.1	89.0

Table 6: Impact of separation and compactness.

Method	HAR	Image cls.
Baseline	84.9	84.2
+ alignment w/o separation	85.0	84.8
+ single-prototype + separation	87.6	88.6
+ multi-prototype + separation	**88.2**	**89.2**

Ablation Study: Ablation studies are conducted on NTU60 and CIFAR-10 using frozen MAMP and MAE encoders, respectively. Best results are shown in **bold**.

Impact of projection layer architecture: Table 4 compares linear and MLP-based projections for aligning frozen features to the TaSS. A linear projection already improves over the baseline, indicating that the TaSS provides a useful target space. The best results are obtained with a pre-normalized three-layer MLP, LN-MLP-3, using a Linear-GELU-Linear-GELU-Linear design, achieving 3.3 and 5.0 percentage-point gains over the baseline on HAR and image classification, respectively. This experiment suggests that moderate nonlinear alignment is beneficial, while deeper projection heads provide no additional gain.

***Impact of the number of prototypes*:** Table 5 reports the effect of varying the number of prototypes used to represent TaSS. Changing the number of prototypes changes the number of modes used to model class-specific variations. In our implementation, the prototypes are obtained by varying the number of clusters in the k-means clustering. We evaluate prototype counts from 2 to 5, and observe the best classification accuracy when 4 prototypes are used.

***Impact of separation and compactness*:** Table 6 shows the step-wise effect of each operation. Alignment without separation provides general semantic-alignment reference and improves over the baseline by 0.1 and 0.6 percentage points on HAR and image classification. Adding separation with a single prototype further improves performance by 2.6 and 3.8 percentage points, showing the benefit of a more separated target space. The best results are obtained with multi-prototype TaSS, reaching 88.2 and 89.2, demonstrating that modeling multiple intra-class modes together with separation produces the most discriminative target space.

***Impact of separation enhancement levels*:** Table 7 reports classification performance under different levels of separation in the TaSS. Here, the pairwise distances between classes are controlled through a concave transformation parameterized by a, where smaller values of a increase inter-class separation and larger values reduce it. As a decreases, the TaSS gradually approaches a uniform geometric distribution. Thus, for a new dataset, a can be selected as the smallest value that maintains high semantic fidelity, assessed using Spearman's ρ, kNN overlap, and ordinal violation rate (OVR). This balances enhanced class separation against the loss of semantic fidelity at excessively small values of a. The best performance of 88.2% and 89.2% is achieved for HAR and image classification, respectively, at $a = 0.2$.

Table 7: Impact of separation enhancement.

Method	HAR	Img.
Baseline	84.9	84.2
Baseline + $a = 1.0$	85.1	84.8
Baseline + $a = 0.3$	87.9	86.5
Baseline + $a = 0.2$	**88.2**	**89.2**
Baseline + $a = 0.1$	87.8	88.4

Table 8: Semantic fidelity evaluation on datasets with high semantic overlap.

Metric	k	NTU120	CUB200
Spearman $\rho \uparrow$	–	0.99	1.00
kNN-Overlap@k $\uparrow$	5	0.99	1.00
	10	0.99	1.00
	100	0.99	1.00
OVR $\downarrow$	–	0.04%	0.00

Table 9: T-to-M and T-to-I retrieval comparison: text encoder semantic-space vs TaSS alignment

T-to-M							T-to-I						
Alignment	$Precision@top\text{-}k\%$			$mAP@top\text{-}k\%$			Alignment	$Precision@top\text{-}k\%$			$mAP@top\text{-}k\%$		
	1%	10%	50%	1%	10%	50%		1%	10%	50%	1%	10%	50%
text encoder	0.97	0.95	0.87	**0.97**	0.93	0.84	text encoder	0.96	0.95	0.88	0.96	0.94	0.85
TaSS	**0.98**	**0.97**	**0.92**	**0.97**	**0.95**	**0.90**	TaSS	**0.98**	**0.97**	**0.94**	**0.97**	**0.96**	**0.93**

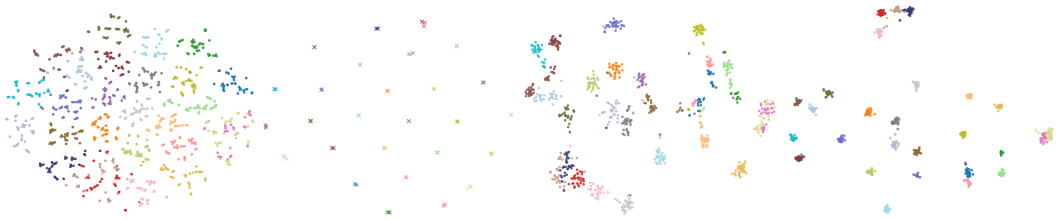
Additional ablations in Appendix section E demonstrate that TaSS is robust to the number of generated text descriptions, the choice of text generating LLM, and the text encoder, with only minor variations in classification accuracy.

Semantic fidelity Validation experiments: Semantic fidelity between the original text-encoder space ρ_E and the TaSS space ρ_M is evaluated using task-agnostic metrics on NTU120 and CUB200. Since the theoretical guarantee relies on a sufficient condition involving the maximum distance approximation error and the minimum target-distance gap, empirical validation is used to assess semantic fidelity in practical fine-grained settings. Following common evaluation practice, three complementary metrics are used: Spearman's ρ for preserving relative class-distance orderings across all class pairs Chen et al. [2024a], Ding et al. [2021], k-nearest-neighbor (kNN) overlap for local geometric consistency between neighborhoods in ρ_E and ρ_M Boggust et al. [2019], and ordinal violation rate (OVR) for the fraction of ρ_E-induced class-pair orderings reversed in ρ_M Suzuki et al. [2021]. Formal definitions are provided in Appendix section C, and results are summarized in table 8. On NTU120, high Spearman correlation and kNN overlap, together with a low OVR of 0.04%, indicate strong semantic-fidelity preservation with only minor ordinal deviations. On CUB200, all metrics reach their maximum values with zero violations, indicating that TaSS preserves the semantic fidelity of the text-encoder space.

Generalization evaluation: We further evaluate TaSS on text-to-motion (T-to-M) and text-to-image (T-to-I) retrieval to assess generalization beyond classification. Class descriptions are used as queries, and retrieval is performed using cosine similarity in the aligned space. USDRL and DINOv3 serve as frozen visual encoders for T-to-M on NTU120 and T-to-I on CUB200, respectively. Performance is measured using $Precision@top\text{-}k\%$ and $mAP@top\text{-}k\%$ Feng et al. [2014].

As shown in table 9, TaSS consistently outperforms the original text-encoder semantic space across both retrieval tasks, with the largest gains at broader retrieval ranges. These results indicate that TaSS preserves semantic structure while improving generalization beyond classification.

Visualization: Figure 2 shows t-SNE visualizations for the 30 hardest CUB200 classes selected using baseline DINOv3 accuracy. Compared with the original text-encoder semantic space (fig. 2a), the constructed TaSS space (fig. 2b) exhibits improved inter-class separation and intra-class compactness while preserving semantic structure. Likewise, compared with frozen DINOv3 embeddings (fig. 2c), TaSS-aligned embeddings (fig. 2d) form tighter and better-separated clusters, demonstrating the discriminative effectiveness of TaSS for image classification.



(a) Original text space (b) TaSS space (c) Baseline image space (d) TaSS aligned image space

Figure 2: t-SNE visualizations of semantic and image embeddings for 30 hard CUB200 classes. a) text-encoder semantic space shows scattered and partially overlapping clusters. b) The constructed TaSS space exhibits improved separation and compactness while retaining semantic fidelity. c) Baseline frozen image embeddings show scattered and class overlaps. d) Image embeddings aligned to TaSS form tighter and better-separated clusters.

8 Conclusion, Limitations, and Future Work

We presented TaSS, a framework for learning a discriminative low-dimensional semantic space for visual classification. TaSS preserves semantic structure while improving inter-class separation and intra-class compactness through a geometry-aware transformation of pretrained text embeddings. Frozen visual features are aligned to TaSS using a lightweight MLP, enabling consistent performance gains across skeleton-based action recognition, video classification, and image classification without fine-tuning pretrained encoders.

The quality of TaSS depends on the coverage and reliability of the underlying text embeddings. In addition, the proposed semantic-fidelity guarantee preserves relative semantic relationships rather than the exact geometry of the original semantic space. The current framework also assumes a fixed class set, motivating future work on incremental and open-vocabulary settings. Extending TaSS to adaptive higher-dimensional target spaces is another promising direction.

References

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL `https://api.semanticscholar.org/CorpusID:268232499`.

Angie Boggust, Brandon Carter, and Arvind Satyanarayan. Embedding comparator: Visualizing differences in global structure and local neighborhoods via small multiples. *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 2019. URL `https://api.semanticscholar.org/CorpusID:209140747`.

Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, and Hujun Bao. Locality sensitive discriminant analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 708–713, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Shan Chen, Jack Gallifant, Mingye Gao, Pedro Moreira, Nikolaj Munch, Ajay Muthukkumar, Arvind Rajan, Jaya Kolluri, Amelia Fiske, Janna Hastings, Hugo J. W. L. Aerts, Brian Anthony, Leo Anthony Celi, William G. La Cava, and Danielle S. Bitterman. Cross-care: Assessing the healthcare implications of pre-training data on language model bias. *Advances in neural information processing systems*, 37 D-ampB:23756–23795, 2024a. URL `https://api.semanticscholar.org/CorpusID:269635575`.

Yang Chen, Jingcai Guo, Song Guo, and Dacheng Tao. Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8721–8730, 2024b. URL `https://api.semanticscholar.org/CorpusID:274131291`.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019. ISBN 978-1-950737-13-0. URL http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DevlinCLT19.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 1556–1568, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:202120592.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Xin Geng, De-Chuan Zhan, and Zhi-Hua Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(6):1098–1107, 2005. doi: 10.1109/TSMCB.2005.850151.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Jussi Karlgren and Pentti Kanerva. Semantics in high-dimensional space. *Frontiers in artificial intelligence*, 4:698809, 2021.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023b.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964. URL https://api.semanticscholar.org/CorpusID:48165675.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

Brian Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4): 287–364, 2013.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, October 2020. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2916873. URL https://doi.org/10.1109/TPAMI.2019.2916873.

Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3d action representation learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10191, 2023.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

David Osowiechi, Mehrdad Noori, Gustavo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ayed, and Christian Desrosiers. Watt: Weight average test time adaptation of clip. *Advances in neural information processing systems*, 37:48015–48044, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

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.

Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Andre Moelle. A comprehensive approach to improving clip-based image retrieval while maintaining joint-embedding alignment. *Information Systems*, 134:102581, 2025.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.*, 8:1027–1061, 2007. URL https://api.semanticscholar.org/CorpusID:9678351.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.

Atsushi Suzuki, Atsushi Nitanda, {Jing Jing} Wang, Linchuan Xu, Kenji Yamanishi, and Marc Cavazza. Generalization error bound for hyperbolic ordinal embedding. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10011–10021, 2021.

Fida Mohammad Thoker, Letian Jiang, Chen Zhao, and Bernard Ghanem. Smile: Infusing spatial and motion semantics in masked video learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8438–8449, 2025.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-ucsd birds-200-2011, 2025. URL https://dx.doi.org/10.21227/fg6w-vh29.

Hongsong Wang, Wanjiang Weng, Junbo Wang, Fang Zhao, Guo-Sen Xie, Xin Geng, and Liang Wang. Foundation model for skeleton-based human action understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *ArXiv*, abs/2109.08472, 2021. URL https://api.semanticscholar.org/CorpusID:237563206.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5022–5030, 2019.

Wenhan Wu, Zhishuai Guo, Chen Chen, Hongfei Xue, and Aidong Lu. Frequency-semantic enhanced variational autoencoder for zero-shot skeleton-based action recognition, 06 2025.

Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10276–10285, October 2023.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *International workshop on deep learning for human activity recognition*, pages 70–84. Springer, 2021.

Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2006.

Tingbing Yan, Wenzheng Zeng, Yang Xiao, Xingyu Tong, Bo Tan, Zhiwen Fang, Zhiguo Cao, and Joey Tianyi Zhou. Crossglg: Llm guides one-shot skeleton-based 3d action recognition in a cross-level manner. *ArXiv*, abs/2403.10082, 2024. URL https://api.semanticscholar.org/CorpusID:268509807.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Anqi Zhu, Qiuhong Ke, Mingming Gong, and James Bailey. Part-aware unified representation of language and skeleton for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18761–18770, June 2024.

Appendix

A Additional Theoretical Analysis

This section provides supplementary derivations supporting the semantic fidelity properties discussed in the main paper.

A.1 Preliminaries and Notation

We recall the key objects introduced in the main text.

- $\mathcal{C} = \{1, \ldots, C\}$: set of semantic classes.
- $\mathcal{P} = \{(c, c') : 1 \leq c < c' \leq C\}$: set of unordered class pairs.
- $\mu_c^E \in \mathbb{S}^{d_E - 1}$: reference semantic centroid of class c.
- $\rho_E(c, c') = \delta_{cc'} = \arccos(\mu_c^E \cdot \mu_{c'}^E)$: semantic dissimilarity of class c, c' in the original text-encoder semantic space.
- $\delta_{\max} = \max_{(c,c') \in \mathcal{P}} \delta_{cc'}$: maximum distance of class pairs in reference space.
- $t_{cc'} = f_a(\delta_{cc'})$ with $f_a(x) = \delta_{\max}(x/\delta_{\max})^a$, $a \in (0, 1]$: target inter-class distances.
- $\mu_c^M \in \mathbb{S}^{d_v - 1}$: learned class centroid in the derived semantic space (TaSS).
- $\rho_M(c, c') = \arccos(\mu_c^M \cdot \mu_{c'}^M)$: learned TaSS metric.

We further recall the quantities used to characterize approximate alignment between learned and target distances:

$$\Delta_{\min} = \min_{\substack{(c,c'),(k,k') \in \mathcal{P} \\ t_{cc'} \neq t_{kk'}}} \left| t_{cc'} - t_{kk'} \right|, \tag{9}$$

which denotes the smallest nonzero gap among target distances, and

$$\varepsilon_\infty = \max_{(c,c') \in \mathcal{P}} \left| \rho_M(c, c') - t_{cc'} \right|, \tag{10}$$

which bounds the worst-case deviation of learned distances from their targets.

Semantic fidelity. As defined in the main paper (Definition 3.1), a derived semantic space $(\mathcal{C}, \rho_M)$ is said to maintain *semantic fidelity* with respect to the reference space $(\mathcal{C}, \rho_E)$ if, for all $(c, c'), (k, k') \in \mathcal{P}$,

$$\rho_E(c, c') < \rho_E(k, k') \;\Rightarrow\; \rho_M(c, c') < \rho_M(k, k'). \tag{11}$$

A.2 Derivation of the Semantic Fidelity Condition

We restate Condition 1 from the main text and provide a formal derivation establishing its sufficiency for semantic fidelity

Condition 1 (Semantic Fidelity Condition (restated))**.** *If*

$$\varepsilon_\infty < \frac{\Delta_{\min}}{2}, \tag{12}$$

then the learned TaSS metric ρ_M does not invert any ordering of inter-class dissimilarities induced by the reference metric ρ_E.

Lemma A.1 (Order preservation under bounded deviation)**.** *Suppose that for all $(c, c') \in \mathcal{P}$,*

$$|\rho_M(c, c') - t_{cc'}| \leq \varepsilon_\infty, \tag{13}$$

and $\varepsilon_\infty < \Delta_{\min}/2$. Then, for any two pairs $(c, c'), (k, k') \in \mathcal{P}$ satisfying

$$t_{cc'} < t_{kk'}, \tag{14}$$

it holds that

$$\rho_M(c, c') < \rho_M(k, k'). \tag{15}$$

Proof. Let $(c, c'), (k, k') \in \mathcal{P}$ with $t_{cc'} < t_{kk'}$. By definition of $\Delta_{\min}$,

$$t_{kk'} - t_{cc'} \geq \Delta_{\min}. \tag{16}$$

Using the triangle inequality,

$$\begin{aligned} \rho_M(k, k') - \rho_M(c, c') &\geq (t_{kk'} - \varepsilon_\infty) - (t_{cc'} + \varepsilon_\infty) && (17)\\ &= t_{kk'} - t_{cc'} - 2\varepsilon_\infty && (18)\\ &\geq \Delta_{\min} - 2\varepsilon_\infty. && (19) \end{aligned}$$

Since $\varepsilon_\infty < \Delta_{\min}/2$, the right-hand side is strictly positive, implying $\rho_M(c, c') < \rho_M(k, k')$. □

Lemma A.2 (Order preservation of target distances)**.** *Let $t_{cc'} = f_a(\rho_E(c, c'))$ with f_a strictly increasing. Then, for any $(c, c'), (k, k') \in \mathcal{P}$,*

$$\rho_E(c, c') < \rho_E(k, k') \;\Rightarrow\; t_{cc'} < t_{kk'}. \tag{20}$$

Proof. Since f_a is strictly increasing, it preserves the ordering of its inputs. Applying f_a to both sides of $\rho_E(c, c') < \rho_E(k, k')$ yields $t_{cc'} < t_{kk'}$. □

Proposition A.3 (Semantic fidelity under bounded deviation)**.** *If $t_{cc'}$ is constructed via a strictly increasing transformation of $\rho_E(c, c')$ and $\varepsilon_\infty < \Delta_{\min}/2$, then the derived semantic space $(\mathcal{C}, \rho_M)$ maintains semantic fidelity with respect to $(\mathcal{C}, \rho_E)$ in the sense of Definition 3.1.*

Proof. For any $(c, c'), (k, k') \in \mathcal{P}$ satisfying $\rho_E(c, c') < \rho_E(k, k')$, Lemma A.2 implies $t_{cc'} < t_{kk'}$. By Lemma A.1, this ordering is preserved by ρ_M under the stated bound on ε_∞, yielding $\rho_M(c, c') < \rho_M(k, k')$. This establishes semantic fidelity. □

A.3 Proofs of Theoretical Properties

We now provide proofs of the theoretical properties stated in the main paper.

Proposition 1 (Semantic fidelity). *If Condition 1 holds, then the derived semantic space $(\mathcal{C}, \rho_M)$ maintains semantic fidelity with respect to $(\mathcal{C}, \rho_E)$.*

Proof. Condition 1 asserts that the maximum deviation between the learned inter-class distances $\rho_M(c, c')$ and their targets $t_{cc'}$ is bounded by $\varepsilon_\infty < \Delta_{\min}/2$. By Proposition A.3 in Appendix A.2, this bound guarantees that no ordering of target distances is inverted by ρ_M. Since the target distances $\{t_{cc'}\}$ preserve the ordering induced by ρ_E via a strictly increasing transformation, it follows that $(\mathcal{C}, \rho_M)$ maintains semantic fidelity with respect to $(\mathcal{C}, \rho_E)$. □

Proposition 2 (Selective separation enhancement). *For concavity parameter $a < 1$, the transformation $t_{cc'} = \delta_{\max}(\delta_{cc'}/\delta_{\max})^a$ expands smaller reference inter-class distances proportionally more than larger ones.*

Proof. Consider the ratio

$$\frac{t_{cc'}}{\delta_{cc'}} = \left(\frac{\delta_{cc'}}{\delta_{\max}}\right)^{a-1} = \delta_{\max}^{1-a}\, \delta_{cc'}^{a-1}.$$

For $a < 1$, we have $a - 1 < 0$. Since $\delta_{cc'} > 0$, the derivative satisfies

$$\frac{d}{d\,\delta_{cc'}}\left(\delta_{\max}^{1-a}\, \delta_{cc'}^{a-1}\right) = \delta_{\max}^{1-a}(a-1)\delta_{cc'}^{a-2} < 0,$$

so the ratio is strictly decreasing in $\delta_{cc'}$. Hence, smaller reference distances receive a larger multiplicative expansion than larger ones, which selectively enhances separation among semantically similar classes. □

Proposition 3 (Effect of prototype-based compactness regularization) *Prototype-based compactness regularization operates only on intra-class dispersion and does not introduce explicit cross-class interactions.*

Proof. The term $\mathcal{L}_{\text{local}}$ depends only on each embedding m_i and its assigned prototype $p^M_{c_i,\pi_i}$, while $\mathcal{L}_{\text{proto}}$ depends only on prototypes $p^M_{c,j}$ and the centroid μ^M_c of the same class. Hence, both terms act only within each class. Therefore, prototype-based compactness regularization reduces intra-class dispersion and introduces no explicit cross-class interactions. □

B Algorithm

Algorithm 1 summarizes the TaSS algorithm for learning the desired semantic space. The method first constructs a target inter-class distance configuration by applying a concave transformation to the distances in the semantic space. A mapping g_θ is then learned to align class-level representations with this target configuration while regularizing within class scatter. The resulting output is the TaSS.

Algorithm 1 TaSS

Require: Text embeddings $E = \{e_i\}_{i=1}^n$, labels $\{c_i\}$, concavity $a \in (0, 1]$, number of prototypes K, steps T

Define $\mathcal{I}_c \leftarrow \{i \mid c_i = c\}$ for each class $c \in \mathcal{C}$
Reference class centroids: compute $\{\mu^E_c\}_{c=1}^C$ from $\{e_i\}$ and $\{\mathcal{I}_c\}$
Reference distances: $\delta_{cc'} \leftarrow \arccos(\mu^E_c \cdot \mu^E_{c'})$ for all $(c, c') \in \mathcal{P}$
$\delta_{\max} \leftarrow \max_{(c,c')\in\mathcal{P}} \delta_{cc'}$
Target distances: $t_{cc'} \leftarrow \delta_{\max}\big(\delta_{cc'}/\delta_{\max}\big)^a$ for all $(c, c') \in \mathcal{P}$
Prototype initialization: for each class c, apply k-means to the original semantic embeddings to obtain K clusters and prototype assignments $\{\pi_i\}$
Initialize $g_\theta : \mathbb{R}^{d_E} \to \mathbb{R}^{d_v}$
for $t = 1, \ldots, T$ **do**
    Project: $m_i \leftarrow g_\theta(e_i)$ for all i
    Class centroids: compute $\{\mu^M_c\}_{c=1}^C$ from $\{m_i\}$ and $\{\mathcal{I}_c\}$
    Prototype centers: compute $\{p^M_{c,j}\}$ from the mapped embeddings $\{m_i\}$ using the fixed prototype assignments $\{\pi_i\}$
    Separation loss: $\mathcal{L}_{\text{sem_geo}} \leftarrow \frac{1}{|\mathcal{P}|}\sum_{(c,c')\in\mathcal{P}} \Big(\arccos(\mu^M_c \cdot \mu^M_{c'}) - t_{cc'}\Big)^2$
    Local compactness loss: $\mathcal{L}_{\text{local}} \leftarrow \frac{1}{n}\sum_{i=1}^n \big\|m_i - p^M_{c_i,\pi_i}\big\|_2^2$
    Prototype compactness loss: $\mathcal{L}_{\text{proto}} \leftarrow \frac{1}{C}\sum_{c=1}^C \frac{1}{K}\sum_{j=1}^K \Big(\big\|p^M_{c,j} - \mu^M_c\big\|_2 - r_{c,j}\Big)^2$
    Update: $\theta \leftarrow \theta - \eta\nabla_\theta\big(\mathcal{L}_{\text{sem_geo}} + \lambda(\mathcal{L}_{\text{local}} + \mathcal{L}_{\text{proto}})\big)$
end for
return TaSS prototypes $\mathcal{P}^M = \{p^M_{c,j}\}_{c=1,j=1}^{C,K}$

C Semantic Validation Experiments

This appendix provides additional details on the evaluation metrics used to empirically validate semantic fidelity, as described in the main paper.

(1) Rank correlation with reference geometry: Spearman's rank correlation is computed between the rankings of inter-class distances induced by the reference semantic metric ρ_E and those induced by the derived semantic metric ρ_M over all class pairs.

- **Spearman rank correlation.** Let $r_E(c, c')$ and $r_M(c, c')$ denote the ranks of $\delta_{cc'}$ and $\rho_M(c, c')$ among all pairs in $\mathcal{P}$. Spearman's ρ is defined as

$$\rho = 1 - \frac{6}{|\mathcal{P}|(|\mathcal{P}|^2 - 1)} \sum_{(c,c')\in\mathcal{P}} \big(r_E(c, c') - r_M(c, c')\big)^2. \tag{21}$$

Higher values of ρ indicate greater similarity between the rankings of inter-class distances induced by ρ_E and ρ_M, providing an empirical proxy for semantic fidelity.

(2) kNN-Overlap@k: kNN overlap evaluates, for each class $c \in \mathcal{C}$, the overlap between the sets of its k nearest neighboring classes under the reference semantic metric ρ_E and under the derived semantic metric ρ_M.

For each class $c \in \mathcal{C}$, let $\mathcal{N}_K^E(c)$ denote the set of the K nearest classes to c under the reference semantic metric ρ_E. Analogously, let $\mathcal{N}_K^M(c)$ denote the set of the K nearest classes to c under the derived semantic metric ρ_M.

$$\text{kNN-Overlap@}K := \frac{1}{C}\sum_{c=1}^{C}\frac{\left|\mathcal{N}_k^E(c) \cap \mathcal{N}_K^M(c)\right|}{K}. \tag{22}$$

As a local comparison of nearest-neighbor relations, kNN overlap@K provides an empirical indicator of how local class neighborhoods induced by ρ_E are reflected under ρ_M.

(3) Ordinal Violation Rate (OVR): The ordinal violation rate quantifies the fraction of ordered class-pair relations induced by ρ_E that are violated under ρ_M. Let,

$$N := \left|\left\{((c,c'),(k,k')) \in \mathcal{P}^2 : \rho_E(c,c') < \rho_E(k,k')\right\}\right|, \tag{23}$$

$$V := \left|\left\{((c,c'),(k,k')) \in \mathcal{P}^2 : \rho_E(c,c') < \rho_E(k,k') \ \wedge \ \rho_M(c,c') \geq \rho_M(k,k')\right\}\right|. \tag{24}$$

Then $\text{OVR} := V/N$.

OVR provides a violation-focused measure of ordinal relations induced by ρ_E that are violated under ρ_M.

D Text Generation and Encoding

Text generation: To capture intra-class variability while preserving class semantics, we generate multiple textual descriptions per class (eg: 150, 300, or 450) using a large language model. The descriptions are obtained by systematically varying meaningful class-specific attributes. In action recognition, for example, these attributes include body symmetry, involved body parts, joint-level involvement, movement amplitude, motion plane, execution style, and repetition. Since the considered tasks differ in semantic structure, separate prompts are designed for human action recognition, image classification, and video classification.

For reproducibility, the prompts used for textual description generation are provided separately below.

1. **Prompt for action description generation.**
   Role: You are a precise biomechanics description engine.
   Action: `pickup something`
   Attributes to vary (with biomechanical plausibility constraints):
   - Side / Symmetry: left, right, both, alternating
   - Body Part Focus: hand, arm, leg, foot, head, trunk, full body
   - Joint-Level Emphasis: shoulder, elbow, wrist, hand, hip, knee, ankle, head/neck, trunk
   - Amplitude: small, medium, large
   - Tempo / Speed: slow, normal, fast, explosive
   - Plane / Direction: sagittal, frontal, transverse
   - Support / Contact Pattern: bilateral, single, alternating, airborne
   - Style / Effort: relaxed, controlled, casual, stiff, powerful, expressive
   - Repetition / Duration: single, repeated, continuous, prolonged
   - Height / Position: low, waist, chest, overhead

Task: Generate exactly 300 distinct natural-language descriptions of the action `pickup something`.

Constraints:

- Each description must incorporate at least three attributes from the list above.
- Each description must consist of 3–5 sentences, where each sentence describes a sub-action or biomechanical phase of the movement.
- All descriptions must be biomechanically realistic, with logically ordered action phases and compatible attribute combinations.
- Attribute combinations should be systematically varied across samples to ensure diversity.
- No descriptions may be duplicated.

Output format: Return an array containing exactly 300 strings.

2. **Prompt for Image classification task**

   Role: You are a precise low-resolution visual appearance description engine specialized for images.

   Object Class: frog **Attributes to vary (with visual plausibility constraints):**

   - Viewpoint / Orientation:front | side | rear | angled | top-down
   - Scale / Object Size:tiny | small | medium
   - Framing / Visibility:centered | off-center | partially cropped | partially visible
   - Occlusion Level:none | mild | heavy
   - Position in Scene:foreground-dominant | background-dominant | blended
   - Background Context:natural | urban | indoor | cluttered | simple
   - Lighting Condition:bright | dim | overcast | shadowed
   - Image Quality / Resolution Artifacts:sharp | slightly blurry | noisy | low-contrast | pixelated
   - Color / Texture Variation:natural colors | muted colors | high contrast

   Task:Generate EXACTLY 300 distinct natural-language image descriptions of the object class “frog”.

   constraints: Each description must:

   - Explicitly mention the object class
   - Incorporate at least 3 different attributes from the list above
   - Each description must be 1–2 sentences only
   - Descriptions must be single-frame only (no temporal progression, no actions)
   - Ensure low-resolution realism consistent with CIFAR images
   - Avoid artistic styles (no sketches, cartoons, paintings)
   - Vary attributes systematically to span intra-class visual variance
   - Do not repeat descriptions

   Output Format: Return a JSON array with EXACTLY 300 strings. No explanations. No extra keys.

3. **Prompt for video classification task Role:** You are a precise video description engine specialized in visual variation for video classification datasets.

   Class Label: <class name>

   Task: Generate EXACTLY 150 distinct natural-language descriptions of short video clips depicting the class “<class name>”.

   Instructions:

   - Infer the likely visual meaning of “<class name>” from the label itself.
   - Keep that meaning fixed across all 150 descriptions.
   - Generate descriptions that reflect realistic intra-class variation while remaining faithful to the same class.

   Rules:

- Every description must explicitly contain the exact class label "<class name>".
- Every description must describe one short video clip only.
- Every description must include at least 4 visible factors chosen from:
  - camera angle or viewpoint
  - framing or crop
  - subject scale
  - temporal crop or motion phase visibility
  - background context
  - lighting condition
  - occlusion or partial visibility
- Vary the descriptions systematically across clip appearance, motion visibility, camera placement, scene type, and video quality.
- Keep the descriptions realistic for natural video datasets.
- Keep each description to 1–2 sentences only.
- Do not mention attribute names explicitly.

Output Format: Return a JSON array with EXACTLY 150 strings. No explanations. No extra keys.

Text encoding. Once the action descriptions are generated, each description is encoded using a pretrained sentence-level text encoder. Specifically, we employ different encoders such as BGE Xiao et al. [2024], CLIP Radford et al. [2021], BERT Devlin et al. [2019].

Each text description is passed through the frozen text encoder, and the final sentence representation is obtained from the output corresponding to the `[CLS]` token. All text embeddings are ℓ_2-normalized prior to subsequent processing. These embeddings serve as the input semantic representations to construct the TaSS.

E Additional Ablation Studies

Additional ablation studies are conducted to further analyze the design choices of the proposed TaSS.

Selection of the mapping function g_θ: To construct the target semantic space $(\mathcal{C}, \rho_M)$ from the original text-encoder semantic space $(\mathcal{C}, \rho_E)$, TaSS employs a learnable mapping function g_θ. To achieve the proposed separation enhancement induced by the targeted distance transformation, the mapping function must provide sufficient capacity to map semantics from the original space to the target space

In this ablation, the effect of the mapping function is studied by varying the capacity of g_θ, ranging from a single linear projection to multi-layer perceptrons with nonlinear activations. The evaluation is performed using the semantic validation experiments described in section C. To isolate the effect of the mapping function, the separation enhancement parameter is fixed to $a = 0.2$ across all experiments.

As shown in table 10, increasing the capacity of the mapping function g_θ improves semantic fidelity. A single linear projection provides insufficient flexibility to realize the targeted distance transformation. Introducing nonlinearity and normalization leads to consistent improvements across all semantic validation metrics, including Spearman correlation and KNN overlap, indicating a more faithful mapping to the target distances. Deeper configurations further enhance semantic fidelity, with a moderately expressive nonlinear mapping proving sufficient.

Sensitivity to the number of descriptions per class: The sensitivity of TaSS to the number of LLM-generated descriptions is evaluated on NTU60 using the MAMP motion encoder, as shown in table 11. Descriptions per class are varied from 50, 150, 300, to 450, with a separate TaSS constructed for each setting and frozen visual features aligned using the MLP. The best performance is obtained with 300 descriptions per class, while results remain relatively stable across settings, indicating that TaSS is not highly sensitive to the number of descriptions.

Sensitivity to the text-generating LLM: As shown in table 12, LLM sensitivity is evaluated on CIFAR-10 with the MAE encoder by varying the text-generating model while keeping the prompt

Table 10: Ablation on the mapping function g_θ. Here, **L** denotes a linear layer, **LN** denotes layer normalization, and **G** denotes the GELU activation. Increasing the capacity of g_θ improves semantic fidelity.

Mapping g_θ	**Spearman**	**KNN-overlap@k**				**OVR**
		5	10	50	100	
L	0.93	0.57	0.63	0.83	0.95	11.4
L → G	0.94	0.58	0.64	0.84	0.95	10.4
L → LN → G	0.98	0.60	0.67	0.91	0.98	5.40
L → LN → G → L	0.99	0.96	0.96	0.99	0.99	0.63
L → LN → G → L → LN	0.99	0.97	0.98	0.99	0.99	0.50
L → LN → G → L → LN → G	0.99	0.90	0.93	0.97	0.97	1.20
L → LN → G → L → LN → G → L	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.04**

Table 11: Impact of the number of text descriptions.

No. of texts	Top-1 (%)
50	87.9
150	88.0
300	**88.2**
450	88.0

Table 12: Sensitivity of TaSS to the text-generating LLM.

LLM used to generate text	Top-1 (%)
ChatGPT-5	**89.2**
Claude	**89.2**
Gemini	89.1

Table 13: Sensitivity of TaSS to the text encoder.

Text encoder used	Top-1 (%)
BGE	**89.2**
CLIP	**89.2**
BERT	89.0

fixed. ChatGPT-5, Google Gemini Anil et al. [2023], and Claude Anthropic [2024] are compared, with Claude and chatGPT-5 achieving the best classification accuracy of 89.2%. Overall, results remain stable across LLMs, indicating that TaSS is not highly sensitive to the text-generating LLM.

Sensitivity to the text encoder: As shown in table 13, text-encoder sensitivity is evaluated on CIFAR-10 with the MAE image encoder using BGE Xiao et al. [2024], CLIP Radford et al. [2021], and BERT Devlin et al. [2019] pretrained text encoders. BGE and CLIP achieve the best performance of 89.2%, while results remain largely consistent, indicating that TaSS is minimally sensitive to the text encoder choice.

F Additional Supervised Baseline Comparisons

F.1 Comparison with Capacity-Matched Supervised MLP Classifier

To investigate whether the improvements of TaSS arise from increased projection-head capacity, we compare TaSS with a supervised MLP classifier of comparable capacity using the same frozen visual encoders. As shown in table 14, TaSS consistently outperforms the capacity-matched supervised MLP across all datasets and encoder architectures, demonstrating the importance of constructing discriminative semantic space. Single seed results are reported

F.2 Comparison with metric-learning baselines

We further extend the metric learning comparison using HAR to NTU60 using the frozen MAMP encoder, with capacity matched projection heads across all methods. As shown in Table 15, TaSS outperforms all four metric-learning baselines by 1.5–2.2 percentage points.

Table 14: Comparison of TaSS with a capacity-matched supervised MLP classifier across HAR, video, and image classification benchmarks. All methods use frozen visual encoders. Top-1 accuracy (%) is reported.

Dataset	Encoder	Capacity-Matched MLP Align.	TaSS MLP Align.
NTU60	USDRL	84.4	**85.6**
	MAMP	85.2	**88.2**
	CTR-GCN	89.8	**90.9**
NTU120	USDRL	75.2	**76.7**
	MAMP	78.7	**81.0**
	CTR-GCN	85.0	**85.9**
HMDB51	SMILE	70.9	**71.8**
	VideoMAEv2	83.6	**84.9**
ARID1.5	SMILE	75.3	**76.9**
	VideoMAEv2	79.2	**80.1**
CIFAR-10	DINOv3	96.9	**98.0**
	MAE	84.5	**89.5**
CUB-200	DINOv3	84.9	**88.4**
	MAE	31.3	**40.2**
Oxford-IIIT Pet	DINOv3	94.3	**95.1**
	MAE	81.2	**84.3**

Table 15: Comparison with metric-learning baselines on NTU60 using the frozen MAMP encoder. All methods use projection heads with comparable capacity. Top-1 accuracy (%) is reported.

Method	Top-1 Accuracy (%)
SupCon	86.1
Multi-Similarity	86.0
Circle Loss	86.7
ArcFace	86.5
TaSS (Ours)	**88.2**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are consistent with the paper's contributions: TaSS learns a task-adaptive low-dimensional semantic space that preserves semantic fidelity while improving class separability. This work supports these claims with theoretical analysis, empirical results across HAR, video, and image classification, ablations, semantic-fidelity validation, and retrieval experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the "Conclusion, Limitations, and Future Work" section. In particular, it discuss that TaSS depends on the coverage and reliability of the underlying text-encoder semantic space. It also clarifies that the semantic-fidelity guarantee preserves relative semantic relationships, rather than the exact geometry of the original text-encoder semantic space.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provide theoretical properties in the main text and provides formal assumptions, definitions, lemmas, propositions, and proofs in Appendix. The appendix includes the semantic-fidelity condition, order-preservation lemmas, and proofs for semantic fidelity, selective separation enhancement, and prototype-based compactness regularization.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the algorithm, datasets, frozen visual encoders, text encoders, prompt-generation procedure, projection design, ablations, and implementation details. Algorithm 1 describes TaSS construction, and the experiments specify datasets, backbones, and evaluation settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code will be made publicly available soon. All datasets used in this work are open-source or publicly available for download. The generated text descriptions are not included in the submission; however, the text-generation prompts and the LLMs used to produce the descriptions are provided in the appendix.

   Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the major experimental settings, including datasets, tasks, frozen encoders, text-generation models, text encoders, projection architecture, and GPU type. It also provides the text-generation prompts, ablation details in Appendix. Additional implementation details, including the optimizer, learning rate, batch size, number of epochs, and random seeds, follow the settings used by the corresponding baseline networks. The train/test splits are also the same as those provided by the original dataset repositories.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports accuracy comparisons, ablation studies, semantic-fidelity metrics, and retrieval metrics across a broad range of tasks, datasets, and encoder architectures. Mean and standard deviation over three independent random seeds are reported only for the Text Encoder MLP Alignment and TaSS MLP Alignment methods in Table 2. The remaining experiments were conducted using a single run, as repeating all experiments across multiple random seeds would require substantial additional computational resources and time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper states that all experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU, but it does not provide memory, runtime per experiment, total compute budget, CPU/storage details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses public benchmark datasets, pretrained encoders, and LLM-generated class descriptions for classification experiments. To the best of the authors' knowledge, the work conforms to the NeurIPS Code of Ethics. No personally identifying data collection, human-subject experiments, or high-risk data release is introduced by the proposed method.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: The paper does not include a dedicated discussion of societal impacts. The current submission focuses on the methodological, theoretical, and empirical aspects of TaSS, which is evaluated on standard public visual-classification benchmarks. Broader societal implications, including potential positive uses and possible risks in sensitive visual-recognition applications, are not explicitly discussed in the paper.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper does not release a high-risk pretrained generative model, scraped dataset, or dual-use model with obvious high misuse potential. The method is an alignment framework for classification using existing benchmarks and pretrained encoders.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper cites existing datasets, pretrained encoders, and models, including NTU RGB+D, HMDB51, ARID, CIFAR-10, CUB200, Oxford-IIIT Pet, CLIP, BERT, BGE, DINOv3, MAE, VideoMAEv2, SMILE, USDRL, and MAMP.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce a new dataset, benchmark, pretrained model, or other standalone asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or new human-subject data collection. It uses existing public datasets and LLM-generated descriptions.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve new crowdsourcing or human-subject experiments, so IRB or equivalent approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are part of the method because class-level textual descriptions are generated using GPT, Gemini, and Claude, then encoded with pretrained text encoders to construct TaSS. The paper describes this usage in the implementation section and provides prompt templates in Appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.