DON'T CARE ABOUT ENCODERS: ENCODER-AGNOSTIC CLASS-INCREMENTAL CROSS-MODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world retrieval systems must continuously evolve to incorporate new semantic classes and upgraded backbone architectures while maintaining retrieval across modalities. However, existing class-incremental learning (CIL) methods often assume fixed encoder architectures, while backward-compatible training typically focuses on unimodal scenarios. To address this challenge, we propose *Don't Care About Encoders* (DCAE), an encoder-agnostic framework designed for class-incremental cross-modal retrieval. DCAE uses lightweight residual adapters on top of frozen backbone encoders to map feature vectors into a shared normalized representation space. To promote geometric stability across modalities and encoder versions, we anchor representations to fixed class prototypes with simplex equiangular tight frame (ETF) geometry. Our method further encourages preservation of backbone relationships through a relational matching loss and mitigates catastrophic forgetting via supervised contrastive distillation and replay. Experimental results on audio-visual retrieval tasks demonstrate that DCAE outperforms the evaluated baselines in final retrieval accuracy, while compatibility matrices assess retrieval against old gallery embeddings without re-encoding the corresponding items. The code will be made available after acceptance.

1 INTRODUCTION

A deployed retrieval system needs to improve without discarding what it has already indexed. As new semantic classes arrive, it should learn to retrieve them while retaining knowledge of earlier ones. This represents a central challenge of class-incremental learning (De Lange et al., 2021; Wang et al., 2025). Over time, stronger pretrained encoders become available, making encoder replacement a natural step toward improving representation quality. However, their new embeddings have to remain comparable with the ones already stored. Adopting a new encoder can require re-encoding the accumulated archive, a cost that grows with gallery size and requires access to the original content. Backward-compatible representation learning addresses this issue by making updated queries directly comparable with previous gallery embeddings (Shen et al., 2020; Biondi et al., 2023a).

Cross-modal retrieval makes this transition even more demanding. Modality encoders can originate from different pretrained models and improve at different rates, motivating replacements at different stages. An upgraded audio encoder, for example, should retrieve visual items indexed before the upgrade, while the system continues learning new classes. Retaining class discrimination is insufficient since representations can remain discriminative while changing their coordinates (*e.g.* , by rotating). Likewise, alignment between the current modality encoders does not ensure compatibility with earlier galleries. Continual adaptation therefore has to preserve both previously acquired knowledge and the comparability of representations across modalities and model versions.

Existing approaches address different aspects of this evolving retrieval problem. Continual cross-modal retrieval studies forgetting and changes in multimodal embedding spaces (Wang et al., 2021); compatible lifelong learning and stationarity-based methods preserve compatibility with earlier representations during sequential updates (Biondi et al., 2023b), including encoder replacements (Biondi et al., 2024; 2026); cross-modal backward-compatible training studies upgrades between

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

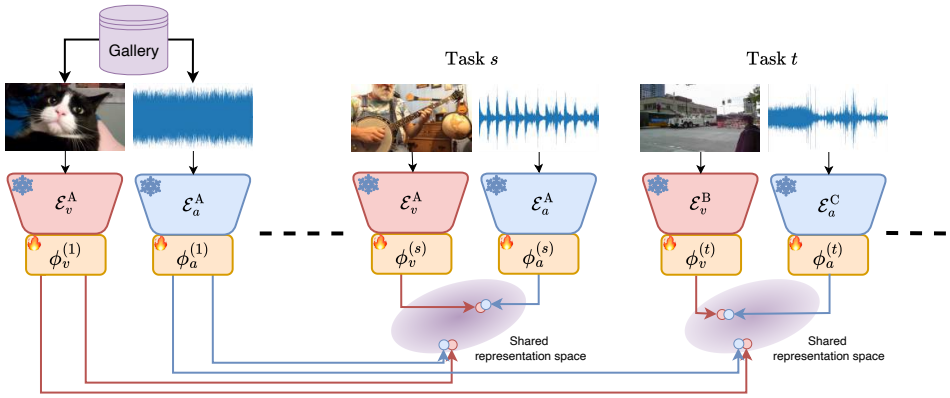


Figure 1: DCAE supports cross-modal retrieval across class increments and encoder replacements. Adapters over frozen backbones map modality-specific features into a shared representation space with fixed class prototypes. Updated queries are compared with gallery embeddings computed by earlier system versions, without re-encoding those gallery items.

vision-language models (Jang & Lim, 2025). We address *class-incremental cross-modal retrieval with encoder replacements*, where heterogeneous modality branches can adopt different pretrained encoders throughout a supervised learning sequence. The objective is to incorporate new classes and exploit improved encoders while retaining cross-modal retrieval against galleries produced by earlier system versions.

We introduce *Don't Care About Encoders* (DCAE), which organizes class-incremental cross-modal retrieval around a persistent geometric reference shared across modalities and encoder versions, as shown in Fig. 1. DCAE fixes a set of class prototypes with Simplex-ETF geometry (Yang et al., 2022; Biondi et al., 2023a) and learns lightweight adapters that map frozen backbone features into a shared representation space. The backbones can change, while the class reference used to train their adapters remains fixed. An encoder replacement therefore changes the mapping into the representation space without redefining its class anchors or re-encoding historical gallery items. This design separates the evolution of the feature extractors from the reference used to maintain cross-modal compatibility throughout the learning sequence.

The central challenge is to align heterogeneous representations with this reference while retaining the useful structure of each backbone and the knowledge acquired in earlier tasks. Our geometric analysis establishes sufficient conditions for backward compatibility in class-conditional expected cosine distance under imperfect prototype alignment. DCAE trains toward this geometry by treating the mapping of backbone features into the shared representation space as a single placement problem. Fixed class prototypes provide common anchors across modalities and encoder versions, relational matching (Park et al., 2019) encourages preservation of the backbone's pairwise distances, and supervised contrastive distillation from stored replay embeddings (Barletti et al., 2021) encourages consistency with historical representations. Together, these objectives balance prototype alignment, preservation of backbone structure, and consistency across tasks, using the same geometric reference throughout the learning sequence.

Our contributions are:

- We formulate and evaluate supervised class-incremental cross-modal retrieval in which modality encoders can be replaced during the learning sequence and updated queries are evaluated against historical gallery representations.
- We propose DCAE, which maps features from heterogeneous modality encoders into a shared representation space anchored by fixed class prototypes through lightweight adapters. Its training objective combines prototype alignment, relational matching of backbone features, and supervised contrastive distillation from replay representations. Our analysis establishes sufficient conditions under which the shared prototype geometry supports backward compatibility in class-conditional expected cosine distance.

- We evaluate DCAE on AVE-CI and VGGSound-CI, with an additional class-based image-text protocol derived from MSCOCO. DCAE improves final retrieval accuracy over the evaluated baselines, while compatibility matrices and ablations characterize its behavior across updates and its dependence on the training objectives and replay budget.

2 Related Work

We address class-incremental cross-modal retrieval with encoder replacements. The objective is to incorporate new classes and exploit improved encoders while keeping updated queries comparable with historical gallery embeddings. We discuss continual learning, backward-compatible representation learning, and multimodal alignment in relation to this objective. Table 4 in Section B summarizes which methods cover cross-modality, class-incremental learning, and encoder replacement.

Continual learning, particularly class-incremental learning, studies how to learn new classes while retaining knowledge of earlier ones, commonly through rehearsal and distillation (Rebuffi et al., 2017; Li & Hoiem, 2017). In multimodal recognition, AV-CIL preserves audio-visual knowledge across class increments (Pian et al., 2023). Retaining class discrimination, however, does not ensure that updated queries remain comparable with earlier gallery embeddings. Wang et al. (2021) study continual cross-modal retrieval by distinguishing training, indexing, and querying. In unimodal retrieval, CL^{2}R preserves representation compatibility during sequential learning (Biondi et al., 2023b), while IAM-CL^{2}R supports asynchronous encoder replacements with improved pretrained models, including changes of architecture (Biondi et al., 2024; 2026). These settings leave open how to preserve cross-modal retrieval against historical galleries when modality encoders are replaced during class-incremental learning. DCAE addresses this joint problem: heterogeneous modality branches can adopt improved pretrained encoders at different stages, while updated queries must remain comparable with gallery embeddings produced by earlier versions and other modalities.

Backward-compatible representation learning makes updated queries directly comparable with previous gallery embeddings and has been studied predominantly in unimodal retrieval (Shen et al., 2020). CoReS uses fixed class geometry to encourage stationary representations across model updates (Biondi et al., 2023a). Extensions beyond unimodal retrieval include the modality-change experiments of BT^2 (Zhou et al., 2023) and, more recently, XBT, which studies upgrades between vision-language models through a learned projection and parameter-efficient adaptation (Jang & Lim, 2025). MKP-Adapter (Byun et al., 2026) further learns backward-compatible mappings from pre-extracted multimodal embeddings while preserving the new model's representation structure, without updating either backbone. Closely related in motivation, OrthAlign (Gupta et al., 2026) aligns independently trained multimodal contrastive spaces through an SVD-based orthogonal map, including transfer across modalities and composition across models. XBT and MKP-Adapter target compatibility with an earlier model's embedding space, while OrthAlign aligns spaces learned by different models. DCAE instead learns mappings into a shared representation space anchored by fixed class prototypes, which remain the same across modalities and encoder versions as new classes arrive and modality encoders are replaced. Each encoder replacement changes the mapping into the representation space without redefining its class anchors or re-encoding historical gallery items.

Multimodal Alignment makes representations from different modalities comparable in a shared embedding space. CLIP and CLAP learn shared embedding spaces for image–text and audio–text retrieval (Radford et al., 2021; Wu* et al., 2023), while ImageBind and LanguageBind align additional modalities using images and language as references, respectively (Girdhar et al., 2023; Zhu et al., 2023). Alignment between the current modality encoders, however, does not ensure compatibility with earlier galleries after continual adaptation or encoder replacement. Fixed class references also have precedents in class-incremental learning: Neural Collapse Terminus uses Simplex-ETF geometry throughout incremental training (Yang et al., 2023). DCAE learns lightweight adapters that map frozen backbone features into a shared representation space, with the same class prototypes used across modalities and encoder versions. The class prototypes, the backbone's pairwise geometry, and the embeddings stored for replay jointly guide the learned representations. Multimodal alignment and compatibility across model versions are thus addressed through the same placement problem against a persistent geometric reference.

3 PROBLEM FORMULATION

Building on the criterion of Jang & Lim (2025, Sec. 3.1), we study class-incremental cross-modal retrieval in which modality-specific encoders may be replaced asynchronously. Representations change as adapters learn new classes and as their frozen backbones are replaced. Our goal is to keep updated queries comparable with historical gallery embeddings across modalities, without re-encoding the gallery.

At task $T^{(t)}$ in the sequence $\mathcal{T} = (T^{(1)}, \dots, T^{(N)})$, training introduces classes $\mathcal{Y}^{(t)}$, with $\mathcal{Y}^{(t)} \cap \mathcal{Y}^{(s)} = \emptyset$ for $t \neq s$. We denote the training minibatch by $\mathcal{B}^{(t)}$ and the replay memory of past examples by $\mathcal{M}$. Let $\mathcal{E}_m^{(t)}$ be the frozen backbone and $\phi_m^{(t)}$ the adapter for modality m at step t. Their composition $f_m^{(t)} = \phi_m^{(t)} \circ \mathcal{E}_m^{(t)}$ produces normalized representations $\boldsymbol{z}_{i,m}^{(t)} = f_m^{(t)}(x_i) \in \mathcal{Z}$, where $\mathcal{Z} \subseteq \mathbb{S}^{d-1}$ denotes the shared normalized representation space. Adapter updates can cause representation drift even with fixed backbones (Yu et al., 2020; Luo et al., 2025); independent encoder replacements introduce another source of change.

Let $\mathcal{Q}^{(t)}$ denote queries of modality m_q encoded by $f_{m_q}^{(t)}$, and $\mathcal{G}^{(s)}$ the gallery of modality m_g encoded by $f_{m_g}^{(s)}$. Query and gallery sets contain different samples from the evaluation classes, including classes not yet observed during training. For $s < t$, compatibility compares the same query samples encoded at steps s and t against fixed gallery embeddings computed at step s.

Definition 3.1 (Cross-modal backward compatibility)**.** *For fixed modalities m_q, m_g and steps $s < t$, the mapping $f_{m_q}^{(t)}$ is backward compatible with $\mathcal{G}^{(s)}$ if, for every evaluated query–gallery pair (x_i, x_j) with labels (y_i, y_j),*

$$D_{\cos}\big(\boldsymbol{z}_{i,q}^{(t)}, \boldsymbol{z}_{j,g}^{(s)}\big) \leq D_{\cos}\big(\boldsymbol{z}_{i,q}^{(s)}, \boldsymbol{z}_{j,g}^{(s)}\big), \qquad y_i = y_j, \tag{1}$$

$$D_{\cos}\big(\boldsymbol{z}_{i,q}^{(t)}, \boldsymbol{z}_{j,g}^{(s)}\big) \geq D_{\cos}\big(\boldsymbol{z}_{i,q}^{(s)}, \boldsymbol{z}_{j,g}^{(s)}\big), \qquad y_i \neq y_j, \tag{2}$$

where $D_{\cos}(\boldsymbol{u}, \boldsymbol{v}) = 1 - \boldsymbol{u}^\top \boldsymbol{v}$ for unit vectors, and q, g identify the query and gallery modalities. Cross-modal backward compatibility requires these conditions for every considered modality pair.

Fixed backbones restrict this setting to continual adapter updates; setting $m_q = m_g$ recovers the unimodal condition. Section 4 studies its class-conditional expectation relaxation and the learning objective that promotes it. Notation is summarized in Section A.

4 A SHARED CLASS GEOMETRY FOR HETEROGENEOUS ENCODERS

DCAE maintains a common representation space as modality-specific adapters learn new classes and their frozen backbones are replaced. The adapters map backbone features toward the same fixed class prototypes, providing a reference shared across modalities and encoder versions. Learning must balance alignment with this reference against preservation of backbone structure and historical representations. We describe this geometric trade-off, use it to motivate the training objective, and characterize compatibility with historical galleries under imperfect alignment.

4.1 SHARED SIMPLEX GEOMETRY

The requirements in Eqs. (1) and (2) constrain every query–gallery pair: even an update that improves retrieval can increase some same-class distances or decrease some different-class distances. Following Biondi et al. (2024; 2026), we study their class-conditional expectation, averaging distances separately for each class pair. For the analysis, the trained encoders and adapters are fixed, and query and gallery items are sampled independently conditional on their labels. Expected cosine distance then depends only on the class means:

$$\mathbb{E}[D_{\cos}(\boldsymbol{z}_{c,q}, \boldsymbol{z}_{j,g})] = 1 - \boldsymbol{m}_{c,q}^\top \boldsymbol{m}_{j,g}, \qquad \boldsymbol{m}_{y,a} = \mathbb{E}[\boldsymbol{z}_{y,a}], \quad a \in \{q, g\}. \tag{3}$$

For a fixed gallery distribution, compatibility in expectation depends on how an update changes the query class means.

To guide these changes consistently across modalities and encoder versions, we give the class means a shared reference. Independently learned simplex arrangements need not share an orientation (Papyan et al., 2020), so following Pernici et al. (2021); Biondi et al. (2023a; 2026), we fix K unit prototypes $\{\boldsymbol{\mu}_y\}_{y=1}^{K} \subset \mathbb{R}^d$ with simplex equiangular tight frame (simplex-ETF) geometry:

$$\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = -\gamma \quad (i \neq j), \qquad \gamma = \frac{1}{K-1}, \qquad 2 \leq K \leq d+1. \tag{4}$$

We allocate enough prototypes for the learning sequence and assign each new class an unused prototype. This assignment remains fixed across modalities and encoder versions, allowing new classes to be introduced without changing earlier class anchors.

Fixing the prototypes does not enforce the learned class means to align exactly with them: each adapter is also trained to preserve backbone structure and information from earlier tasks. We account for this imperfect alignment by decomposing each mean into its prototype component and an orthogonal offset, suppressing domain and step indices:

$$\boldsymbol{m}_y = \alpha_y \boldsymbol{\mu}_y + \boldsymbol{b}_y, \qquad \alpha_y = \boldsymbol{\mu}_y^\top \boldsymbol{m}_y, \qquad \boldsymbol{b}_y \perp \boldsymbol{\mu}_y. \tag{5}$$

Here α_y measures prototype alignment and $\boldsymbol{b}_y$ is the orthogonal offset of the class mean. Together with the within-class variation $\rho_y = \mathbb{E}\|\boldsymbol{z} - \boldsymbol{m}_y\|_2^2$, they satisfy the unit-normalization identity

$$\alpha_y^2 + \|\boldsymbol{b}_y\|_2^2 + \rho_y = 1. \tag{6}$$

For $\alpha_y \geq 0$, increasing alignment leaves less room for $\|\boldsymbol{b}_y\|_2^2 + \rho_y$. An aligned mean ($\boldsymbol{b}_y = \mathbf{0}$) can still retain within-class variation ρ_y, whereas $\alpha_y = 1$ collapses the class onto its prototype. This trade-off motivates promoting alignment while preserving backbone structure and, across updates, encouraging consistency with historical class means.

4.2 Learning the Shared Representation

A trainable classification head supervises class predictions through cross-entropy. To encourage alignment with the fixed class prototypes, we additionally minimize a prototype-alignment loss, the expected squared distance between the adapter outputs and their corresponding class prototype:

$$\mathcal{L}_{\mathrm{ETF}} = \mathbb{E}_{(x_m,y)} \|\boldsymbol{z}_m^{(t)} - \boldsymbol{\mu}_y\|_2^2, \qquad \ell_y = \mathbb{E}\|\boldsymbol{z}_y - \boldsymbol{\mu}_y\|_2^2 = 2 - 2\alpha_y. \tag{7}$$

The identity $\ell_y = 2 - 2\alpha_y$ directly connects this objective to the analysis in Proposition 4.1: reducing the loss for class y increases its prototype alignment α_y.

However, prototype attraction alone favors classwise collapse and can remove distinctions between samples of the same class (Laurent et al., 2024). We therefore complement alignment with a relational matching loss, penalizing changes in backbone pairwise distances (Park et al., 2019):

$$\mathcal{L}_{\mathrm{Rel}} = \frac{1}{n} \sum_m \left\| D(\mathrm{Z}_m^{(t)}) - D(\mathrm{H}_m^{(t)}) \right\|_{\mathrm{F}}^2. \tag{8}$$

Here the rows of $\mathrm{H}_m^{(t)} \in \mathbb{R}^{n \times d_{m,t}}$ and $\mathrm{Z}_m^{(t)} \in \mathbb{R}^{n \times d}$ contain normalized backbone features and adapter outputs, respectively, and D is the pairwise Euclidean distance matrix. This loss discourages distortion of both within-class and between-class relationships. In particular, $\mathbb{E}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 = 2\rho_y$ for independent class-y samples, so penalizing distortion of these distances encourages preservation of within-class variation ρ_y (Section C.1).

Alignment and structure preservation must also extend across tasks: learning new classes need not improve alignment for earlier ones, and encoder replacements introduce further changes in the source features. Replay memory $\mathcal{M}$ therefore stores past examples, labels, and previous adapter outputs $\widetilde{\boldsymbol{z}}_{i,m}$, providing historical supervision without retaining earlier encoders. Following Barletti et al. (2021), we use these stored outputs for supervised contrastive distillation:

$$\mathcal{L}_{\mathrm{SupCon},m} = -\mathbb{E}_{i,p} \log \frac{\exp((\boldsymbol{z}_{i,m}^{(t)})^\top \widetilde{\boldsymbol{z}}_{p,m}/\tau)}{\sum_{j \neq i} \exp((\boldsymbol{z}_{i,m}^{(t)})^\top \widetilde{\boldsymbol{z}}_{j,m}/\tau)}. \tag{9}$$

Here $\tau > 0$ is the temperature, and the expectation averages uniformly over replay anchors with another same-class example, then over their same-class positives $p \neq i$; the denominator includes

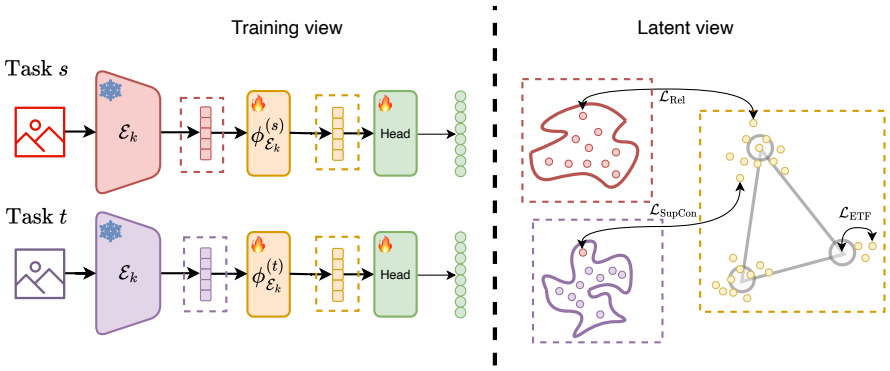


Figure 2: DCAE across tasks $s < t$. Adapters map frozen backbone features into a shared representation space. Fixed simplex-ETF prototypes supply class anchors, while relational matching and supervised contrastive distillation provide structural and temporal supervision. Only the prototypes have fixed simplex geometry; the feature distributions are learned.

all other replay examples in the minibatch. These positive-pair similarities encourage current representations to remain consistent with stored same-class targets, which may originate from different steps. Their relation to changes in class means is given in Section C.2.

With $\mathcal{L}_{\text{SupCon}}$ denoting distillation aggregated over modalities, the complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{ETF}}\mathcal{L}_{\text{ETF}} + \lambda_{\text{Rel}}\mathcal{L}_{\text{Rel}} + \lambda_{\text{SupCon}}\mathcal{L}_{\text{SupCon}}. \tag{10}$$

The loss weights balance prototype alignment, backbone structure, and historical consistency (Fig. 2). At inference, only the backbone and adapter are used; the classification head is discarded.

4.3 Compatibility with Historical Galleries

The objective encourages alignment with shared prototypes while retaining structure and historical information. We now relate the resulting class geometry to compatibility: for each query–gallery modality pair, the following proposition gives sufficient conditions under which alignment gains outweigh the effects of changes in orthogonal offsets. Its proof is given in Section C.3.

Proposition 4.1 (Cross-modal backward compatibility in class-conditional expectation)**.** *Fix steps $s < t$ and query and gallery modalities q, g. For each class represented in the query and gallery evaluation sets, let $\boldsymbol{m}_c^{(r)}$ be its query mean at $r \in \{s, t\}$ and $\bar{\boldsymbol{m}}_j$ its gallery mean at s, using the same query input distribution at both steps. Decompose these means as in Eq.* (5)*, using barred coefficients and offsets for the gallery. Assume:*

(i) All domains share the unit prototypes in Eq. (4)*.*

(ii) Representations have unit norm, and query and gallery items are independent conditional on their labels and the fixed trained encoders and adapters.

(iii) Query alignment does not decrease, $\Delta\alpha_c = \alpha_c^{(t)} - \alpha_c^{(s)} \geq 0$, and gallery alignment is nonnegative, $\bar{\alpha}_j \geq 0$, for all evaluated classes.

Define $\Delta\boldsymbol{b}_c = \boldsymbol{b}_c^{(t)} - \boldsymbol{b}_c^{(s)}$, $\bar{\varepsilon}_j = \|\bar{\boldsymbol{b}}_j\|_2$, $\bar{\rho}_j = 1 - \|\bar{\boldsymbol{m}}_j\|_2^2$, and

$$d_{cj}^r = \mathbb{E}[D_{\cos}(\boldsymbol{z}_{c,q}^{(r)}, \boldsymbol{z}_{j,g}^{(s)})], \qquad \Delta d_{cj} = d_{cj}^t - d_{cj}^s.$$

Then

$$\Delta d_{cc} \leq -\Delta\alpha_c\bar{\alpha}_c + \|\Delta\boldsymbol{b}_c\|_2\bar{\varepsilon}_c, \tag{11}$$

$$\Delta d_{cj} \geq \gamma\Delta\alpha_c\bar{\alpha}_j - \Delta\alpha_c\bar{\varepsilon}_j - \|\Delta\boldsymbol{b}_c\|_2\sqrt{1-\bar{\rho}_j}, \qquad j \neq c. \tag{12}$$

Compatibility in class-conditional expectation holds if the upper bound in Eq. (11) *is nonpositive for every class and the lower bound in Eq.* (12) *is nonnegative for every different-class pair. If all relevant query and gallery offsets vanish, assumption (iii) suffices.*

The bounds show why a shared simplex is useful: increasing query alignment α_c contributes to decreasing same-class distances and increasing different-class distances, because the corresponding prototype inner products are 1 and $-\gamma$. Orthogonal offsets can counteract these gains, but their effect depends on the class pair. For the same class, query offset changes interact only with the gallery offset and have no effect when that gallery mean is aligned; for different classes, offsets can also project onto other prototype directions.

These effects explain the complementary roles of the losses: $\mathcal{L}_{\mathrm{ETF}}$ promotes alignment α_y, $\mathcal{L}_{\mathrm{Rel}}$ encourages preservation of within-class variation ρ_y, and $\mathcal{L}_{\mathrm{SupCon}}$ encourages consistency of historical class means, including their orthogonal offsets $\boldsymbol{b}_y$. The proposition concerns expected distances rather than retrieval accuracy. Complementing this analysis, Section D examines the combined contribution of the additional losses and replay through prototype alignment across tasks.

5 Experiments

Datasets. We run the main experiments on two, well established, class-incremental audio-visual datasets, proposed by Pian et al. (2023). The first is based on AVE (Tian et al., 2018), while the second one is based on VGGSound (Chen et al., 2020). To evaluate DCAE on another modality pair, we additionally consider a class-based image–text protocol derived from MSCOCO 2014 (Lin et al., 2014) More details about the datasets can be found in Section E.1.

Encoder replacements. The audio-visual protocol begins with the image encoder from CLIP and the audio encoder from CLAP. We subsequently replace these branches with SigLIP and M2D-CLAP, respectively, and finally with the image and audio encoders from ImageBind. For AVE-CI, we apply the following encoder replacements: SigLIP replaces CLIP during Task 2 and onwards, M2D-CLAP replaces CLAP from Task 3, and ImageBind replaces both models for the final task. On VGGSound-CI, the same changes are applied at Task 4, Task 7 and Task 9, respectively. On MSCOCO, we apply a different strategy due to the nature of the dataset. We construct the class-incremental split by dividing all 80 classes among 4 tasks, resulting in 20 classes per task. We retain images with a single annotated category to construct disjoint class-incremental tasks. The first two tasks use the CLIP image encoder and the CLAP text encoder, so that the initial branches come from different pretrained models. For the last two tasks, we replace them with the corresponding ImageBind encoders. As in the audio-visual protocol, adapters are trained on top of the frozen encoders in all tasks.

Adapter architecture. We introduce a lightweight residual projection as our adapters on top of the frozen backbone encoders. Their main goal is to align the raw, heterogeneous embeddings to our shared geometrically structured representation space $\mathcal{Z}$. Concretely, given an input embedding $\boldsymbol{h}_m^{(t)} \in \mathbb{R}^{d_{m,t}}$, the adapter applies a two-layer MLP, with GELU activations, omitting the bias terms, and a residual skip connection, mapping $\boldsymbol{h}_m^{(t)}$ to $\boldsymbol{z}_m^{(t)} \in \mathcal{Z}$. We apply ℓ_2-normalization after the residual addition to constrain adapter outputs to the unit hypersphere. The fixed class prototypes have simplex-ETF geometry, while the learned representations are encouraged to align with them. We mathematically describe our adapter structure using Eq. (13). For backbone feature dimensions no larger than the shared dimension, we zero-pad the input features $\boldsymbol{h}_m^{(t)}$ obtaining $\bar{\boldsymbol{h}}_m^{(t)}$ of the required dimension. The residual connection allows the adapter to express its transformation as a correction to the padded backbone features, while relational matching discourages distortion of their pairwise distances.

$$r_m^{(t)}(\bar{\boldsymbol{h}}_m^{(t)}) = W_2\,\sigma(W_1\bar{\boldsymbol{h}}_m^{(t)}), \quad \boldsymbol{z}_m^{(t)} = \frac{\bar{\boldsymbol{h}}_m^{(t)} + r_m^{(t)}(\bar{\boldsymbol{h}}_m^{(t)})}{\|\bar{\boldsymbol{h}}_m^{(t)} + r_m^{(t)}(\bar{\boldsymbol{h}}_m^{(t)})\|_2}, \tag{13}$$

where $r_m^{(t)}(\cdot)$ denotes the residual MLP, W_i represents the weight matrix associated with the ith layer and σ is the activation function.

Baselines. We compare our method with representative methods for multimodal alignment, backward compatibility, and class-incremental learning, such as: OrthAlign (Gupta et al., 2026), XBT (Jang & Lim, 2025), AV-CIL (Pian et al., 2023), NC Terminus (Yang et al., 2023), and MKP-Adapt (Byun et al., 2026). We adapted all methods to our scenario, *i.e.* class-incremental cross-modal

Table 1: Comparison between previous methods and our new approach. Best results, excluding oracles, are highlighted in **bold**.

Dataset	Method	Audio-to-Image				Image-to-Audio			
		FA ↑	AA ↑	AC ↑	ACA ↑	FA ↑	AA ↑	AC ↑	ACA ↑
AVE-CI	OrthAlign	$20.41_{\pm 0.00}$	$22.87_{\pm 0.00}$	$0.50_{\pm 0.00}$	$15.03_{\pm 0.00}$	$24.91_{\pm 0.00}$	$26.93_{\pm 0.00}$	$0.67_{\pm 0.00}$	$22.88_{\pm 0.00}$
	XBT	$6.91_{\pm 9.06}$	$16.03_{\pm 0.99}$	$0.11_{\pm 0.19}$	$1.48_{\pm 2.55}$	$10.52_{\pm 3.96}$	$12.06_{\pm 2.31}$	$0.39_{\pm 0.10}$	$4.57_{\pm 0.65}$
	AV-CIL	$54.19_{\pm 2.90}$	$24.43_{\pm 3.49}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$59.42_{\pm 2.10}$	$25.78_{\pm 2.35}$	$0.17_{\pm 0.00}$	$6.34_{\pm 0.95}$
	NC Terminus	$68.38_{\pm 4.30}$	$30.74_{\pm 5.61}$	$0.39_{\pm 0.25}$	$12.62_{\pm 10.53}$	$73.08_{\pm 5.04}$	$40.22_{\pm 7.45}$	$0.61_{\pm 0.34}$	$23.55_{\pm 18.43}$
	MKP-Adapt	$12.14_{\pm 1.00}$	$8.84_{\pm 2.24}$	$0.17_{\pm 0.17}$	$2.06_{\pm 2.46}$	$13.63_{\pm 4.60}$	$15.41_{\pm 3.61}$	$0.28_{\pm 0.10}$	$6.28_{\pm 1.05}$
	DCAE (ours)	$\mathbf{77.47}_{\pm 0.62}$	$\mathbf{41.57}_{\pm 0.71}$	$\mathbf{0.83}_{\pm 0.16}$	$\mathbf{30.95}_{\pm 4.61}$	$\mathbf{84.29}_{\pm 1.34}$	$\mathbf{48.52}_{\pm 0.47}$	$\mathbf{0.94}_{\pm 0.09}$	$\mathbf{41.48}_{\pm 4.90}$
VGG-CI	OrthAlign	$14.19_{\pm 0.00}$	$16.33_{\pm 0.00}$	$0.53_{\pm 0.00}$	$11.15_{\pm 0.00}$	$15.68_{\pm 0.00}$	$16.50_{\pm 0.00}$	$0.44_{\pm 0.00}$	$8.94_{\pm 0.00}$
	XBT	$0.87_{\pm 0.81}$	$7.46_{\pm 5.44}$	$0.09_{\pm 0.15}$	$0.15_{\pm 0.27}$	$1.28_{\pm 0.46}$	$4.69_{\pm 3.21}$	$0.33_{\pm 0.06}$	$1.72_{\pm 1.23}$
	AV-CIL	$30.37_{\pm 2.40}$	$10.40_{\pm 1.39}$	$0.24_{\pm 0.09}$	$3.02_{\pm 1.37}$	$31.23_{\pm 2.43}$	$10.91_{\pm 1.50}$	$0.20_{\pm 0.12}$	$2.63_{\pm 1.36}$
	NC Terminus	$64.36_{\pm 7.28}$	$25.66_{\pm 3.33}$	$0.47_{\pm 0.06}$	$16.87_{\pm 3.55}$	$61.33_{\pm 4.92}$	$26.37_{\pm 2.26}$	$0.54_{\pm 0.19}$	$17.86_{\pm 5.54}$
	MKP-Adapt	$1.00_{\pm 0.01}$	$1.38_{\pm 0.34}$	$0.22_{\pm 0.06}$	$0.33_{\pm 0.17}$	$1.70_{\pm 1.33}$	$2.41_{\pm 0.83}$	$0.21_{\pm 0.22}$	$0.94_{\pm 1.00}$
	DCAE (ours)	$\mathbf{67.35}_{\pm 0.69}$	$\mathbf{28.90}_{\pm 0.92}$	$\mathbf{1.00}_{\pm 0.00}$	$\mathbf{28.45}_{\pm 0.91}$	$\mathbf{64.25}_{\pm 0.98}$	$\mathbf{30.23}_{\pm 0.73}$	$\mathbf{0.93}_{\pm 0.08}$	$\mathbf{26.68}_{\pm 3.69}$

retrieval with encoder replacements, as there are no available benchmarks for our proposed task. All implementations are done using PyTorch.All methods use the same task splits, encoder replacement schedule, and replay memory size as DCAE, with method-specific hyperparameters follow the original implementations. OrthAlign's method closely relates to our approach, as they also aim at aligning the embeddings of older backbones to newer ones, but focuses solely on the modalities of contrastive models and using orthogonal alignment. XBT introduces a two-step approach, *i.e.* pretraining on text and fine-tuning on the other modality, using a contrastive objective, to obtain backward compatibility in a multi-modal scenario. AV-CIL addresses audio–visual class-incremental classification rather than retrieval, and thus focuses on an efficient fusion method rather than on aligning the embeddings. NC Terminus represents a solution for class incremental learning. We introduced it as a Simplex-based baseline to showcase our method improvements over simply using the Simplex-ETF structure in the same scenario. Finally, MKP-Adapt proposes the training of an adapter for backwards compatibility using a multimodal large language model as query encoder and a multi-level knowledge preservation loss. For more details on implementations, please refer to Section E.

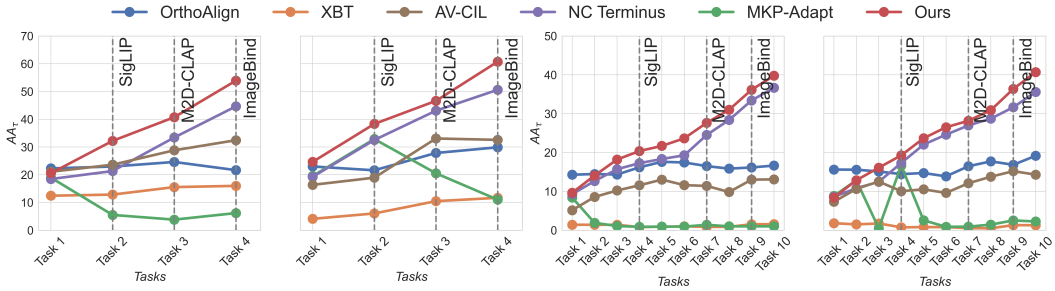
Evaluation protocol. The performance of all evaluated methods is measured following prior works (Ramanujan et al., 2022; Zhou et al., 2023; Biondi et al., 2023a). We adopt two different scenarios for our evaluation: audio-to-image and image-to-audio retrieval, using the query and gallery set, formally defined in Section 3. We measure retrieval performance using the top-1 Cumulative Matching Characteristic (CMC@1), the proportion of queries whose highest-ranked gallery item has the same class label with items ranked by cosine distance. We construct several compatibility matrices to measure retrieval performance across query and gallery system versions. For all tasks, we encode the samples from the full query and gallery sets, not just the ones corresponding to the current task. A detailed view of all the retrieval compatibility matrices is available in Sections H to J. We also report several aggregation metrics, which are extracted from the more comprehensive view of the compatibility matrix. More specifically, we report the Final Accuracy (FA), *i.e.* the CMC@1 for the last task, Average Accuracy (AA) and Average Compatibility (AC) proposed by Biondi et al. (2023a), while ACA is an extension, first proposed by Biondi et al. (2026). AC and ACA summarize empirical retrieval compatibility across system versions. These metrics assess retrieval performance rather than satisfaction of the pairwise distance inequalities or their class-conditional expectation relaxation. Additionally, we report AA_τ, representing the AA computed over the first τ steps.

5.1 Comparative study

We present a comparative analysis of our proposed method (DCAE) against the baseline models. The audio-visual tasks are presented in Table 1, while the visual-language ones are in Table 2. Results in Tables 1 and 2 are averaged over three independent runs. Several ablation studies are available in Section F, including for each component, their associated λ from Eq. (10), replay buffer size, and batch size for the $\mathcal{L}_{\text{SupCon}}$. Moreover, the Appendix contains the performances of the encoders alone, both in CIL and in a non-CIL scenario, in Section H for the AVE dataset, in Section I for VGGSound one and in Section J for MSCOCO.

Table 2: Comparison on MSCOCO 2014 CIL. Best results are highlighted in **bold.**

Dataset	Method	Image-to-Text				Text-to-Image			
		FA ↑	AA ↑	AC ↑	ACA ↑	FA ↑	AA ↑	AC ↑	ACA ↑
MSCOCO-CI	OrthoAlign	$42.36_{\pm 0.00}$	$\mathbf{35.52}_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$38.62_{\pm 0.00}$	$\mathbf{34.31}_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
	XBT	$36.30_{\pm 4.62}$	$31.71_{\pm 1.04}$	$0.11_{\pm 0.10}$	$3.59_{\pm 3.11}$	$29.53_{\pm 7.01}$	$27.90_{\pm 2.02}$	$0.33_{\pm 0.17}$	$10.23_{\pm 4.05}$
	AV-CIL	$41.56_{\pm 1.50}$	$17.28_{\pm 0.24}$	$0.28_{\pm 0.19}$	$6.96_{\pm 2.26}$	$43.75_{\pm 0.70}$	$16.40_{\pm 0.65}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
	NC Terminus	$43.55_{\pm 1.75}$	$16.78_{\pm 1.04}$	$0.56_{\pm 0.35}$	$10.14_{\pm 3.27}$	$35.79_{\pm 15.19}$	$16.19_{\pm 4.14}$	$0.22_{\pm 0.25}$	$5.48_{\pm 6.02}$
	MKP-Adapt	$17.25_{\pm 1.13}$	$9.12_{\pm 0.41}$	$0.22_{\pm 0.19}$	$1.62_{\pm 1.40}$	$9.29_{\pm 4.09}$	$5.50_{\pm 0.87}$	$0.11_{\pm 0.19}$	$0.47_{\pm 0.81}$
	DCAE (ours)	$\mathbf{50.88}_{\pm 0.22}$	$27.59_{\pm 1.15}$	$\mathbf{1.00}_{\pm 0.00}$	$\mathbf{27.24}_{\pm 1.64}$	$\mathbf{52.24}_{\pm 1.05}$	$27.87_{\pm 0.99}$	$\mathbf{0.94}_{\pm 0.10}$	$\mathbf{25.21}_{\pm 1.98}$



(a) A2I on AVE-CI (b) I2A on AVE-CI (c) A2I on VGGSound-CI (d) I2A on VGGSound-CI

Figure 3: Plots of AA_τ for each individual task. The vertical lines show the task at which the encoder is replaced, accompanied by the name of the new encoder.

For the audio-visual datasets, DCAE obtains the highest FA, AA, AC, and ACA among the evaluated methods in both retrieval directions. The strongest baseline in final accuracy is NC Terminus, which also uses fixed simplex-ETF class prototypes. On AVE-CI, DCAE improves mean final accuracy over NC Terminus by 9.09 percentage points for audio-to-image retrieval and 11.21 points for image-to-audio retrieval. On VGGSound-CI, DCAE achieves AC values of 1.00 and 0.93, compared with the highest baseline values of 0.53 and 0.54, respectively. These results show improvements in both final retrieval accuracy and empirical retrieval compatibility under the evaluated encoder-replacement schedules. Moreover, Fig. 3 shows the AA_τ obtained at each task, for the first run of each method. On MSCOCO-CI, DCAE achieves the highest FA, AC and ACA for both Image-to-Text and Text-to-Image. However, OrthoAlign achieves a highest AA in both direction, but without retaining any compatibility, showing decreasing accuracy as tasks progresses in our more relaxed scenario.

6 Conclusion

We introduced DCAE, a framework for class-incremental cross-modal retrieval with encoder replacements. Lightweight residual adapters map frozen backbone features into a shared normalized representation space, where fixed simplex-ETF class prototypes provide a common reference across modalities and encoder versions. Relational matching encourages preservation of backbone relationships, while supervised contrastive distillation and replay encourage consistency with historical representations. Our analysis gives sufficient conditions for backward compatibility in class-conditional expected cosine distance under imperfect prototype alignment. Experiments on AVE-CI, VGGSound-CI, and a class-based MSCOCO protocol show higher final retrieval accuracy than the evaluated baselines, while compatibility matrices assess retrieval against previous galleries. The limitations and future directions are discussed in Section G. These findings support organizing evolving encoders around a persistent geometric reference to maintain comparability with previously indexed representations.

AI USE STATEMENT

In this work, we used generative AI tools to help organize the literature, experimental results, tables and figures; revise the manuscript text and provide text suggestions. The authors have evaluated all AI generated content and take responsibility for the final method, experiments, results and the manuscript.

ETHICS STATEMENT

This work does not involve human subjects, user studies, crowdsourcing, or the collection of new data from individuals. All experiments use publicly available datasets. The authors declare no conflicts of interest and have read and adhered to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Sections 3 to 5 and Section E specify all the details needed for reproducing our method. More specifically, Section 3 formulates the problem, Section 5 provides the proposed method's training objectives while Section 5 gives the main details about our method's setting. Finally, Section E provides more in-depth descriptions for the datasets, model used and adaptation of the baselines.

REFERENCES

Tommaso Barletti, Niccoló Biondi, Federico Pernici, Matteo Bruni, and Alberto Del Bimbo. Contrastive supervised distillation for continual representation learning. In *ICIAP*, 2021.

Niccolo Biondi, Federico Pernici, Matteo Bruni, and Alberto Del Bimbo. CoReS: Compatible representations via stationarity. In *IEEE TPAMI*, 2023a.

Niccolo Biondi, Federico Pernici, Matteo Bruni, Daniele Mugnai, and Alberto Del Bimbo. Cl2r: Compatible lifelong learning representations. In *ACM TOMM*, 2023b.

Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. Stationary representations: Optimally approximating compatibility and implications for improved model replacements. In *CVPR*, 2024.

Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. A Stationary (and Therefore Compatible) Representation is All You Need. In *TPAMI*, 2026.

Jaeseok Byun, Gukyeong Kwon, Han-Kai Hsu, Meher Gitika Karumuri, Zhikang Zhang, Hao Yang, and Davide Modolo. Multi-modal Knowledge Preserving Adapter for Embedding Backward Compatibility. In *ECCV*, 2026.

Haoran Chen, Houze Xu, Micah Goldblum, Daoguo Dong, and Zuxuan Wu. Preserving Cross-Modal Consistency for CLIP-based Class-Incremental Learning. In *ICCV*, 2025.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A Large-scale Audio-Visual Dataset. In *ICASSP*, 2020.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. In *IEEE TPAMI*, 2021.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Sharut Gupta, Sanyam Kansal, Stefanie Jegelka, Phillip Isola, and Vikas Garg. Canonicalizing Multimodal Contrastive Representation Learning. In *arXiv preprint 2602.17584*, 2026.

Young Kyun Jang and Ser-nam Lim. Towards Cross-modal Backward-compatible Representation Learning for Vision-Language Models. In *arXiv preprint 2405.14715*, 2025.

Thomas Laurent, James H. von Brecht, and Xavier Bresson. Feature Collapse. In *ICLR*, 2024.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *IEEE TPAMI*, 2017.

Yu Liang, Yufeng Zhang, Shiliang Zhang, Yaowei Wang, Sheng Xiao, Rong Xiao, and Xiaoyu Wang. MixBCT: Towards Self-Adapting Backward-Compatible Training. In *arXiv preprint 2308.06948*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

Mao-Lin Luo, Zi-Hao Zhou, Tong Wei, and Min-Ling Zhang. LADA: Scalable Label-Specific CLIP Adapter for Continual Learning. In *ICML*, 2025.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. In *PNAS*, 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *CVPR*, 2019.

Federico Pernici, Matteo Bruni, Claudio Baecchi, and Alberto Del Bimbo. Regular polytope networks. In *IEEE TNNLS*, 2021.

Weiguo Pian, Shentong Mo, Yunhui Guo, and Yapeng Tian. Audio-visual class-incremental learning. In *ICCV*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Forward compatible training for large-scale embedding retrieval systems. In *CVPR*, 2022.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards Backward-Compatible Representation Learning. In *CVPR*, 2020.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-Visual Event Localization in Unconstrained Videos. In *ECCV*, 2018.

Kai Wang, Luis Herranz, and Joost van de Weijer. Continual learning in cross-modal retrieval. In *CVPR workshop CLVISIOn*, 2021.

Yan Wang, Da-Wei Zhou, and Han-Jia Ye. Integrating Task-Specific and Universal Adapters for Pre-Trained Model-based Class-Incremental Learning. In *CVPR*, 2025.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *NeurIPS*, 2022.

Yibo Yang, Haobo Yuan, Xiangtai Li, Jianlong Wu, Lefei Zhang, Zhouchen Lin, Philip Torr, Dacheng Tao, and Bernard Ghanem. Neural Collapse Terminus: A Unified Solution for Class Incremental Learning and Its Variants. In *ICLR extension*, 2023.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic Drift Compensation for Class-Incremental Learning. In *CVPR*, 2020.

Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, and Ser-Nam Lim. BT^2: Backward-compatible Training with Basis Transformation. In *ICCV*, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2023.

A NOTATIONS

The notations used in this work are summarized in Table 3.

Table 3: Summary of notation and definitions.

Symbol	Description
x, x_m	Input and modality-specific input.
y, c, j	Class labels in the geometric analysis.
m, m_q, m_g	Modality, query modality, and gallery modality.
q, g, a	Query/gallery identifiers; $a \in \{q, g\}$.
$\mathcal{T}, T^{(t)}, N$	Task sequence, task t, and number of tasks.
t, s, r, τ	Update steps, with $s < t$ and $r \in \{s, t\}$.
$\mathcal{Y}^{(t)}$	Classes introduced for training at task t.
$\mathcal{B}^{(t)}, \mathcal{M}$	Training minibatch and replay memory.
$\mathcal{E}_m^{(t)}, \phi_m^{(t)}$	Frozen backbone and adapter for modality m at step t.
$f_m^{(t)}$	Representation mapping $\phi_m^{(t)}(\mathcal{E}_m^{(t)})$.
$\mathcal{Z} \subseteq \mathbb{S}^{d-1}$	Shared representation space on the unit sphere.
$d, d_{m,t}$	Shared-space and backbone feature dimensions.
$\boldsymbol{z}_{i,m}^{(t)}, \boldsymbol{h}, \tilde{\boldsymbol{h}}$	Normalized adapter output, backbone feature and zero-padded backbone feature.
$\mathcal{Q}^{(t)}, \mathcal{G}^{(s)}$	Query and gallery sets encoded at t and s.
$\mathbb{E}[\cdot]$	Expectation operator.
$D_{\cos}$	Cosine distance: $1 - \boldsymbol{u}^\top \boldsymbol{v}$ for unit vectors.
$\mathrm{H}_m^{(t)}, \mathrm{Z}_m^{(t)}, n$	Backbone/output feature matrices and row count.
$D(\cdot)$	Pairwise Euclidean distance matrix.
$K, \boldsymbol{\mu}_y$	Prototype capacity and fixed unit class prototype.
γ	Simplex separation: $\boldsymbol{\mu}_c^\top \boldsymbol{\mu}_j = -\gamma = -1/(K-1)$ for $c \neq j$.
$\boldsymbol{m}_y, \boldsymbol{m}_{y,a}$	Class mean and domain-specific class mean.
$\alpha_y, \boldsymbol{b}_y$	Prototype alignment and orthogonal mean offset.
ρ_y	Within-class variation: $1 - \|\boldsymbol{m}_y\|_2^2$.
ℓ_y	Class-conditional prototype-alignment loss: $2 - 2\alpha_y$.
$\boldsymbol{m}_c^{(r)}, \bar{\boldsymbol{m}}_j$	Query mean at r. Bars mark fixed gallery statistics.
$\Delta\alpha_c, \Delta\boldsymbol{b}_c$	Query alignment and offset changes from s to t.
$\bar{\varepsilon}_j$	Gallery offset magnitude: $\|\bar{\boldsymbol{b}}_j\|_2$.
$d_{cj}^r, \Delta d_{cj}$	Expected query–gallery distance and its change.
$\widetilde{\boldsymbol{z}}_{i,m}, \widetilde{\boldsymbol{m}}_y$	Stored replay feature and class mean; tildes mark stored-target statistics.
A_y	Expected positive similarity: $(\boldsymbol{m}_y^{(t)})^\top \widetilde{\boldsymbol{m}}_y$.
d_z, d_h	Same-class distances in adapter/backbone spaces.
$\eta_y, \rho_y^{\mathcal{H}}$	Root mean squared relational distortion and backbone within-class variation.
$N_{c,q}^{(t)}, \widehat{\boldsymbol{m}}_{c,q}^{(t)}$	Class sample count and empirical query mean.
$\widehat{R}_{c,q}^{(t)}$	Mean resultant length $\|\widehat{\boldsymbol{m}}_{c,q}^{(t)}\|_2$; hats mark empirical statistics.
$\mathcal{L}_{\mathrm{ETF}}, \mathcal{L}_{\mathrm{Rel}}, \mathcal{L}_{\mathrm{SupCon}}$	Prototype-alignment, relational matching, and supervised contrastive distillation losses.
$\lambda_{\mathrm{ETF}}, \lambda_{\mathrm{Rel}}, \lambda_{\mathrm{SupCon}}$	Auxiliary loss weights.
$\mathcal{L}_{\mathrm{CE}}, \mathcal{L}$	Cross-entropy and complete objective.
$r_m^t(\cdot)$	Residual MLP
W_i, σ	Layer-i weight matrix and activation function.

B Comparison with Existing Methods

Table 4 positions DCAE with respect to the methods discussed in Section 2 along the three axes of our setting. *Cross Modality* indicates that the method operates on data from more than one modality, *e.g.* audio–visual or image–text. *Class Incremental* indicates sequential learning over disjoint class sets, with at most a bounded replay memory of earlier tasks. *Encoder Replacement* indicates that the encoder can be replaced by a different pretrained model, possibly with a different architecture, while embeddings computed by the previous model remain in use; sequential fine-tuning of the same encoder does not satisfy this criterion. Each method listed in Table 4 covers at most two of the three axes; DCAE addresses all three jointly.

Table 4: Comparison with previous methods along the three axes of our setting. ✗ denotes an absent criterion and ✓ a present one. Methods evaluated in Section 5 are marked with †.

Method	Cross Modality	Class Incremental	Encoder Replacement
Wang et al. (2021)	✓	✓	✗
AV-CIL† (Pian et al., 2023)	✓	✓	✗
DMC-OT (Chen et al., 2025)	✓	✓	✗
NC Terminus† (Yang et al., 2023)	✗	✓	✗
CL^{2}R (Biondi et al., 2023b)	✗	✓	✗
IAM-CL^{2}R (Biondi et al., 2024; 2026)	✗	✓	✓
BCT (Shen et al., 2020)	✗	✗	✓
MixBCT (Liang et al., 2024)	✗	✗	✓
XBT† (Jang & Lim, 2025)	✓	✗	✓
MKP-Adapter† (Byun et al., 2026)	✓	✗	✓
OrthAlign† (Gupta et al., 2026)	✓	✗	✓
DCAE (ours)	✓	✓	✓

C Compatibility Analysis

C.1 Relational Distortion and Within-Class Variation

For two independent inputs from class y, let $d_z = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$ and $d_h = \|\boldsymbol{h}_i - \boldsymbol{h}_j\|_2$, using paired, normalized adapter and backbone representations for each input. Define the population relational distortion $\eta_y^2 = \mathbb{E}[(d_z - d_h)^2]$ and the backbone within-class variation $\rho_y^{\mathcal{H}}$. For independent unit-normalized samples,

$$\mathbb{E}[d_z^2] = 2\mathbb{E}\|\boldsymbol{z}\|_2^2 - 2\|\mathbb{E}[\boldsymbol{z}]\|_2^2 = 2\rho_y, \qquad \mathbb{E}[d_h^2] = 2\rho_y^{\mathcal{H}}.$$

Consequently,

$$\begin{aligned} |\rho_y - \rho_y^{\mathcal{H}}| &= \frac{1}{2}\,|\mathbb{E}[(d_z - d_h)(d_z + d_h)]| \\ &\le \frac{1}{2}\eta_y \sqrt{\mathbb{E}[(d_z + d_h)^2]} \\ &\le \eta_y \sqrt{\rho_y + \rho_y^{\mathcal{H}}} \le \sqrt{2}\,\eta_y. \end{aligned} \tag{14}$$

The inequalities use Cauchy–Schwarz, $(a+b)^2 \le 2a^2 + 2b^2$, and $\rho_y, \rho_y^{\mathcal{H}} \le 1$. Thus small within-class relational distortion limits changes in within-class variation. The minibatch loss estimates distortion over its sampled pairs.

C.2 Supervised Contrastive Distillation and Changes in Class Means

Suppress the modality index and let $\widetilde{\boldsymbol{m}}_y = \mathbb{E}[\widetilde{\boldsymbol{z}}_y]$ be the stored target mean for class y. Under independent class-conditional sampling, the expected positive similarity is $A_y = \mathbb{E}[(\boldsymbol{z}_y^{(t)})^\top \widetilde{\boldsymbol{z}}_y] =$

$(\boldsymbol{m}_y^{(t)})^\top \widetilde{\boldsymbol{m}}_y$. With tildes also denoting stored-target alignment, offset, and variation,

$$\begin{aligned}\|\boldsymbol{m}_y^{(t)} - \widetilde{\boldsymbol{m}}_y\|_2^2 &= 2 - \rho_y^{(t)} - \widetilde{\rho}_y - 2A_y \\ &= (\alpha_y^{(t)} - \widetilde{\alpha}_y)^2 + \|\boldsymbol{b}_y^{(t)} - \widetilde{\boldsymbol{b}}_y\|_2^2. \end{aligned} \tag{15}$$

The first equality uses $\|\boldsymbol{m}\|_2^2 = 1 - \rho$; the second follows from orthogonality to the shared prototype. For fixed $\rho_y^{(t)}$ and $\widetilde{\rho}_y$, increasing A_y reduces mean displacement. This explains the role of positive similarities without identifying the full contrastive objective with mean-distance minimization. Targets from several steps define a stored mixture, rather than the representation of one particular step.

C.3 Proof of the Compatibility Proposition

We prove Proposition 4.1 for fixed steps $s < t$ and modalities q, g. Query means at both steps are defined on the same class-conditional input distributions, and the gallery distribution and its step-s mapping remain fixed. Evaluated classes need not have been used for training at step s. Each evaluated class is associated with the same prototype at both steps, including when assessing the earlier mapping before that class was introduced for training.

Proof. Conditional independence gives

$$d_{cj}^r = 1 - \mathbb{E}[\boldsymbol{z}_{c,q}^{(r)}]^\top \mathbb{E}[\boldsymbol{z}_{j,g}^{(s)}] = 1 - (\boldsymbol{m}_c^{(r)})^\top \bar{\boldsymbol{m}}_j, \qquad r \in \{s, t\}. \tag{16}$$

Since the gallery mean is unchanged,

$$\begin{aligned}\Delta d_{cj} &= -(\boldsymbol{m}_c^{(t)} - \boldsymbol{m}_c^{(s)})^\top \bar{\boldsymbol{m}}_j \\ &= -\Delta\alpha_c \bar{\alpha}_j \boldsymbol{\mu}_c^\top \boldsymbol{\mu}_j - \Delta\alpha_c \boldsymbol{\mu}_c^\top \bar{\boldsymbol{b}}_j - \Delta\boldsymbol{b}_c^\top \bar{\boldsymbol{m}}_j. \end{aligned} \tag{17}$$

The three terms are the prototype-alignment contribution and its interactions with the gallery and query offsets.

For $j = c$, both $\bar{\boldsymbol{b}}_c$ and $\Delta\boldsymbol{b}_c = \boldsymbol{b}_c^{(t)} - \boldsymbol{b}_c^{(s)}$ are orthogonal to $\boldsymbol{\mu}_c$. Therefore,

$$\begin{aligned}\Delta d_{cc} &= -\Delta\alpha_c \bar{\alpha}_c - \Delta\boldsymbol{b}_c^\top \bar{\boldsymbol{b}}_c \\ &\le -\Delta\alpha_c \bar{\alpha}_c + \|\Delta\boldsymbol{b}_c\|_2 \bar{\varepsilon}_c, \end{aligned} \tag{18}$$

by Cauchy–Schwarz, proving Eq. (11).

For $j \neq c$, the simplex inner product is $-\gamma$, so

$$\Delta d_{cj} = \gamma\Delta\alpha_c \bar{\alpha}_j - \Delta\alpha_c \boldsymbol{\mu}_c^\top \bar{\boldsymbol{b}}_j - \Delta\boldsymbol{b}_c^\top \bar{\boldsymbol{m}}_j. \tag{19}$$

Since $\Delta\alpha_c \ge 0$ and $\|\boldsymbol{\mu}_c\|_2 = 1$,

$$-\Delta\alpha_c \boldsymbol{\mu}_c^\top \bar{\boldsymbol{b}}_j \ge -\Delta\alpha_c |\boldsymbol{\mu}_c^\top \bar{\boldsymbol{b}}_j| \ge -\Delta\alpha_c \bar{\varepsilon}_j, \tag{20}$$

$$-\Delta\boldsymbol{b}_c^\top \bar{\boldsymbol{m}}_j \ge -\|\Delta\boldsymbol{b}_c\|_2 \|\bar{\boldsymbol{m}}_j\|_2. \tag{21}$$

Unit normalization gives $\bar{\rho}_j = \mathbb{E}\|\boldsymbol{z}_{j,g}^{(s)} - \bar{\boldsymbol{m}}_j\|_2^2 = 1 - \|\bar{\boldsymbol{m}}_j\|_2^2$. Combining these inequalities proves Eq. (12).

A nonpositive same-class upper bound implies $d_{cc}^t \le d_{cc}^s$; a nonnegative different-class lower bound implies $d_{cj}^t \ge d_{cj}^s$. Applying these conditions to every evaluated class pair establishes compatibility in class-conditional expectation.

When all relevant query offsets at s, t and gallery offsets at s vanish, Eq. (17) reduces to

$$\Delta d_{cc} = -\Delta\alpha_c \bar{\alpha}_c \le 0, \qquad \Delta d_{cj} = \gamma\Delta\alpha_c \bar{\alpha}_j \ge 0 \quad (j \neq c), \tag{22}$$

under assumption (iii). □

Finite galleries. The proof also applies to a fixed finite gallery when its class means are empirical means and items are sampled uniformly and independently of the query within each class. In particular, averaging over all pairs from two fixed class sets gives the mean-inner-product identity exactly. This empirical statement does not require the estimated means to equal the corresponding population means.

Table 5: Class concentration $\widehat{R}_{c,q}$ for the seven Task 1 classes of AVE-CI. Original denotes normalized backbone embeddings. Δ_{2-1} and $\Delta_{2-\mathrm{orig}}$ compare Task 2 with Task 1 and Original, respectively.

Class ID	1	2	3	4	5	6	7
Original	0.8513	0.7350	0.8688	0.8262	0.8013	0.7313	0.8074
DCAE (full objective)							
Task 1	0.8894	0.7809	0.8925	0.8045	0.8184	0.7667	0.7250
Task 2	0.9436	0.8886	0.9119	0.8845	0.9053	0.8648	0.8461
Δ_{2-1}	$+0.0542$	$+0.1077$	$+0.0194$	$+0.0800$	$+0.0869$	$+0.0981$	$+0.1211$
$\Delta_{2-\mathrm{orig}}$	$+0.0923$	$+0.1536$	$+0.0431$	$+0.0583$	$+0.1040$	$+0.1335$	$+0.0387$
CE + ETF							
Task 1	0.9610	0.8858	0.8525	0.9062	0.8261	0.8763	0.8188
Task 2	0.8361	0.7050	0.6783	0.7964	0.7351	0.6906	0.7247
Δ_{2-1}	-0.1249	-0.1808	-0.1742	-0.1098	-0.0910	-0.1857	-0.0941
$\Delta_{2-\mathrm{orig}}$	-0.0152	-0.0300	-0.1905	-0.0298	-0.0662	-0.0407	-0.0827

Prototype capacity. For $\Delta\alpha_c > 0$, the different-class sufficient condition is equivalently

$$\bar{\varepsilon}_j + \frac{\|\Delta \boldsymbol{b}_c\|_2}{\Delta\alpha_c}\sqrt{1-\bar{\rho}_j} \le \frac{\bar{\alpha}_j}{K-1}. \tag{23}$$

With alignment and offset statistics fixed, larger prototype capacity K reduces the offset effects that this bound can tolerate. This concerns the sufficient bound: the signed interactions in Eq. (17) can still yield compatibility when the bound is inconclusive. For $\Delta\alpha_c = 0$, use Eq. (12) directly.

D Effect of the Training Objective on Class Geometry

We compare full DCAE with a baseline trained using $\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{ETF}}\mathcal{L}_{\mathrm{ETF}}$ on the seven classes introduced in Task 1 of AVE-CI. The baseline disables $\mathcal{L}_{\mathrm{Rel}}$ and $\mathcal{L}_{\mathrm{SupCon}}$ while retaining classification supervision and prototype attraction. Our aim is to characterize how the training objective affects the concentration of previously learned classes across updates.

For normalized query representations, we compute the mean resultant length

$$\widehat{R}_{c,q}^{(t)} = \left\|\widehat{\boldsymbol{m}}_{c,q}^{(t)}\right\|_2, \qquad \widehat{\boldsymbol{m}}_{c,q}^{(t)} = \frac{1}{N_{c,q}^{(t)}}\sum_{n=1}^{N_{c,q}^{(t)}} \boldsymbol{z}_{n,c,q}^{(t)}, \tag{24}$$

where $N_{c,q}^{(t)}$ is the number of evaluated query samples of class c. The individual representations have unit norm; the empirical class mean is not subsequently normalized. With within-class variation computed on the same samples using divisor $N_{c,q}^{(t)}$, we have $\widehat{\rho}_{c,q}^{(t)} = 1 - (\widehat{R}_{c,q}^{(t)})^2$. Thus, larger $\widehat{R}$ indicates greater class concentration and lower within-class variation.

Applying the class-mean decomposition in Eq. (5) to the empirical means gives

$$(\widehat{R}_{c,q}^{(t)})^2 = (\widehat{\alpha}_{c,q}^{(t)})^2 + \left\|\widehat{\boldsymbol{b}}_{c,q}^{(t)}\right\|_2^2. \tag{25}$$

Whereas $\widehat{\alpha}$ measures the projection onto the assigned prototype, $\widehat{R}$ combines the magnitudes of the prototype-aligned and orthogonal components of the class mean. We therefore use $\widehat{R}$ to characterize overall class concentration independently of the direction of the class mean. This statistic also permits comparison with the original backbone embeddings, which are not trained against the assigned prototypes.

From Task 1 to Task 2, full DCAE increases concentration for all seven classes, whereas the CE + ETF baseline decreases it for every class (Table 5). The full objective also exceeds the Original values at Task 2, while the baseline falls below them. This comparison supports the combined contribution of $\mathcal{L}_{\mathrm{Rel}}$ and $\mathcal{L}_{\mathrm{SupCon}}$ to maintaining concentrated representations of previously learned

classes across this transition, beyond classification supervision and prototype attraction. It does not isolate the two losses individually.

These results characterize changes in class concentration. An increase in $\widehat{R}$ does not by itself imply improved prototype alignment or compatibility with old galleries, and therefore does not establish the alignment and offset conditions of Proposition 4.1.

E Additional Implementation Details

E.1 Detailed dataset description

We run several experiments on two audio-visual datasets, namely AVE (Tian et al., 2018) and VGGSound (Chen et al., 2020). We used the class-incremental (CI) versions for both datasets, selecting the number of tasks and classes per task as proposed in Pian et al. (2023). The AVE-CI dataset consists of 4,134 videos, each of 10 seconds, for an approximate total of 11.5 hours of audio-visual data. It contains 28 events, divided into 4 tasks, with 7 classes per task. The VGGSound-CI is a subset of the original VGGSound, with 100 classes divided into 10 tasks, resulting in 10 classes for each task. It consists of almost 60,000 10-second videos, from which, per class, 50 samples are selected for validation and another 50 for testing. For both datasets, we consider the training split for our class-incremental learning, from which we partition 10% as validation during the training phase. The original validation and test splits are used to define the query and gallery sets, respectively.

For the vision-language experiments, we used MSCOCO (Lin et al., 2014). which contains 82,783 training and 40,504 validation images, each paired with five human-written captions and annotated with object instances from 80 categories. Since most COCO images contain objects from several categories, we keep only the images whose instance annotations belong to a single category. The resulting dataset dataset keeps all 80 classes, divided into 4 tasks with 20 classes per task, ordered by the original category IDs. We kept the first caption of each sample for the language branch.

E.2 Detailed overview of our proposed method

For all experiments, we employ the following frozen pre-trained backbones: for CLIP, we use the `CLIP-ViT-B32`, while the CLAP version is `CLAP-HTSAT-Fused`. For SigLIP, we use the `SigLIP2-B32-256`, and M2D-CLAP is set to the most recent version at the time of writing, which is `M2D-CLAP-ViT-B-80x1001p16x16p16kpBpTI-2025`. Finally, we use the `ImageBind-Huge` version of the model. For all backbone encoders, we precompute the embeddings and save them to disk in order to reduce the computation overhead of training the adapters.

For each task training, we use the Adam optimizer with a learning rate of $1e^{-4}$ and train for up to 250 epochs, using an early stopping mechanism if the validation loss plateaus for more than 25 epochs. We use a batch size of 64 for all of our experiments, which are conducted on a single NVIDIA RTX A6000 with 48 GB of VRAM. A full run on AVE-CI takes around 1.25 hours to complete, while on VGGSound-CI it takes approximately 3 hours to complete.

The objective function for our method is weighted as described in Section F.2. Additionally, we define 3 oracles, composed of encoder combinations, *i.e.* CLIP and CLAP, SigLIP and MD2-CLAP and finnaly, ImageBind. For the oracles, we do not impose constraints and let them train only for the classification task, taking the embeddings right after they pass through the adapter. The results for the oracles are available in Table 6.

E.3 Other methods adaptation

Because none of the previous works is intended for our specific task, we dedicate this section to describing the necessary adaptations for each one. For OrthAlign Gupta et al. (2026), we use the training set to compute the parameters of their proposed ProcrustesAligner and apply them to obtain the final embeddings for the query and the gallery before doing the retrieval protocol. For XBT Jang & Lim (2025), we first pre-trained using their proposed method, *i.e.* LoRA on the backbone encoder with additional soft prompts, on the `Youtube-8M_AudioSetCaps`[1] texts. After this

[1] `https://huggingface.co/datasets/baijs/AudioSetCaps/tree/main/Dataset`

Table 6: Oracle results against our proposed method. All values are determined from a single run.

Dataset	Method	Audio-to-Image				Image-to-Audio			
		FA ↑	AA ↑	AC ↑	ACA ↑	FA ↑	AA ↑	AC ↑	ACA ↑
AVE-CI	Oracle 1	29.20	21.89	0.67	14.96	48.69	31.56	0.33	10.53
	Oracle 2	66.79	44.62	1.00	42.19	69.07	43.91	1.00	40.38
	Oracle 3	72.48	58.31	1.00	57.31	76.61	54.91	1.00	52.78
	DCAE (ours)	76.95	42.29	0.83	31.04	85.83	48.42	0.83	35.83
VGG-CI	Oracle 1	36.35	23.29	0.47	7.66	31.40	19.54	0.33	7.63
	Oracle 2	47.75	29.56	0.49	11.90	40.77	24.67	0.33	5.07
	Oracle 3	39.28	30.31	0.60	16.34	46.70	31.98	0.49	13.88
	DCAE (ours)	67.03	28.90	1.00	27.50	63.17	29.39	0.84	22.54

Table 7: Ablation studies for our method. We use ✗ to denote a removed component, while ✓ denotes a kept one. A2I denotes Audio-to-Image retrieval, while I2A stands for Image-to-Audio retrieval.

Components					A2I	I2A
RB	$\mathcal{L}_{\text{ETF}}$	$\mathcal{L}_{\text{Rel}}$	$\mathcal{L}_{\text{SupCon}}$	RS	FA ↑	FA ↑
✗	✓	✓	✓	10%	71.05	84.00
✓	✗	✓	✓	10%	70.56	81.60
✓	✓	✗	✓	10%	76.00	82.97
✓	✓	✓	✗	10%	76.43	82.03
✓	✓	✓	✓	1%	69.56	77.55
✓	✓	✓	✓	5%	76.09	82.39
✓	✓	✓	✓	10%	76.95	85.83

step, we apply the same CIL setup, where we allow their specific adapter to train as described in their paper,*i.e.* one epoch where only the normalization layers can be updated. For AV-CIL Pian et al. (2023), the only modification is represented by the change of task, as the model was initially intended only for audio-visual classification. While we followed their training strategy, we evaluated the trained model for retrieval. We follow NC Terminus's fixed simplex-ETF terminus over the full label space, the miss-alignment loss and the distillation loss. Moreover, we set their required 20 samples per class replay memory. For MKP-Adapt, we follow their one adapter implementation, keeping their original architecture and specific losses.

F Ablations

F.1 Main components ablation

The ablation studies, presented in Table 7, combined with the proposed method description from Section 4, provide a detailed look into the effects of each component from the proposed framework.

Removing the residual block reduces FA from 76.95 to 71.05 for audio-to-image retrieval and from 85.83 to 84.00 for image-to-audio retrieval. Removing relational matching reduces FA by 0.95 and 2.86 percentage points, respectively. Among the individual loss removals in Table 7, removing prototype alignment produces the largest decrease in both directions: 6.39 and 4.23 percentage points. Removing supervised contrastive distillation also lowers FA, while reducing the replay budget from 10% to 1% produces larger decreases. These results support the contributions of the training components to final retrieval accuracy. Additional experiments with a fixed number of replay samples are available in Section F.3.

Table 8: Ablation studies on the effect of the λ_{ETF}. All reported values represent the FA score.

λ_{ETF}	0.1	1	10	25	50	100	250	500	1000	5000
Audio-to-Image	69.86	64.14	72.32	75.21	74.33	<u>76.68</u>	**76.79**	75.65	74.47	74.49
Image-to-Audio	82.45	83.75	81.11	82.09	<u>84.36</u>	**85.55**	83.58	82.85	82.76	82.67

Table 9: Ablation studies on the effect of the λ_{Rel}. All reported values represent the FA score.

λ_{Rel}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Audio-to-Image	76.16	76.46	76.04	<u>76.86</u>	**76.95**	76.09	75.79	76.40	76.52	76.52
Image-to-Audio	84.64	84.43	84.77	83.76	79.47	84.00	84.79	83.67	**85.71**	<u>85.34</u>

F.2 Loss Weights Ablation

We evaluate the impact of the different components of our loss function by conducting ablation studies for each of the weighting coefficients. The results can be consulted in Tables 8 to 10.

The loss-weight ablations characterize the balance between prototype alignment, preservation of backbone relationships, and supervised contrastive distillation. The ablated value range for λ_{ETF} is selected to be high because we empirically observe that the value for $\mathcal{L}_{\text{ETF}}$ rapidly decreases during the first few steps. For all the main experiments, because we have the best, or second-best, value, we set $\lambda_{\text{ETF}} = 100$. For λ_{Rel}, we observe overall stability, with small improvements when increasing the value for this parameter. As such, following the results obtained in the I2A task, we choose $\lambda_{\text{Rel}} = 1.0$ for the rest of the experiments. Because of the massive drop in accuracy when reducing the replay memory size, we additionally studied the effects of the weight for our distillation loss, presented in Table 10, where higher weights cause performance degradation, most likely due to the model not being plastic enough to learn new classes effectively.

F.3 Dependency on buffer size

In real scenarios, maintaining a large replay buffer of raw data creates several problems for a large-scale retrieval system. The main issues are related to storage and privacy constraints, especially under stricter regulations. To assess dependence on retained data, we vary the replay budget for both our method and XBT, available in Table 11. Despite our drop in performance, which is expected in a replay-free scenario, our method still achieves the best results, some of which are higher than the replay-based method compared in the main manuscript, Table 1. Moreover, from these results, we can see that keeping only 100 samples as a fixed budget memory reduces the catastrophic forgetting problem.

We would like to propose two possible research directions starting from this concern, and present them as possible future works: (i) storing vertex-adjacent synthetic features rather than samples, which the fixed geometry of the ETF makes more tractable, and (ii) the redesign of $\mathcal{L}_{\text{SupCon}}$ with a prototype-anchored distillation term, which would remove the retention requirements at the cost of intra-class structure

F.4 Batch-Size Sensitivity of Supervised Contrastive Distillation

We evaluate sensitivity to training batch size because the supervised contrastive distillation loss depends on the availability of same-class positives and on the composition of the replay examples in each minibatch. Table 12 reports the final accuracy (FA) and average compatibility (AC) for

Table 10: Ablation studies on the effect of the λ_{SupCon}. All reported values represent the FA score.

λ_{SupCon}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Audio-to-Image	**76.95**	70.38	<u>73.39</u>	70.92	71.72	72.35	66.45	67.97	64.36	55.69
Image-to-Audio	**85.83**	77.89	82.51	<u>83.36</u>	78.13	64.26	76.43	74.94	72.60	68.89

Table 11: Effect of the replay buffer size on AVE-CI. We compare exemplar-free training, fixed-count buffers, and buffers sized proportionally to the task data. † denotes the setting reported in Table 1.

Replay buffer	Audio-to-Image		Image-to-Audio	
	FA ↑	AC ↑	FA ↑	AC ↑
		XBT		
None	18.82	0.33	17.21	0.50
100 examples	18.36	0.33	16.58	0.33
500 examples	17.68	0.50	19.39	0.50
10% of data†	17.37	1.00	15.09	0.83
		DCAE (ours)		
None	38.11	0.66	43.70	0.50
100 examples	73.36	0.83	76.82	1.00
500 examples	76.03	1.00	84.52	1.00
10% of data†	76.95	0.83	85.83	0.83

Table 12: Effect of the training batch size on AVE-CI. The last row uses gradient accumulation (ga) to reach an effective batch size of 256. † denotes the setting used in all other experiments.

Batch size	Audio-to-Image		Image-to-Audio	
	FA ↑	AC ↑	FA ↑	AC ↑
4	75.09	0.83	84.09	1.00
16	76.58	1.00	83.85	0.66
32	75.18	1.00	87.77	0.83
64†	76.95	0.83	85.83	0.83

several batch sizes. The reported results show no monotonic decrease in FA as batch size is reduced. However, fixed class prototypes do not make the distillation objective independent of minibatch composition.

G LIMITATIONS

While the DCAE framework demonstrates advancements under class-incremental cross-modal retrieval with encoder replacements, some limitations remain, offering possible future directions for research.

Firstly, we want to address the capacity of the Simplex-ETF prior, which is the core of our geometric stability. However, since the dimensionality of it directly affects the number of equidistant prototypes, *i.e.* $K \leq d + 1$ in scenarios involving extreme scale incremental learning, *e.g.* thousands of classes, this parameter needs to increase proportionally to still maintain its property.

The dependence on backbone discriminability is another limitation of our work. DCAE adopts an approach based on keeping the backbone encoder frozen and training only lightweight adapters. Freezing the backbones avoids backbone fine-tuning, but the adapters depend on the information retained in the pretrained features, ultimately limiting the system's performance. This is especially the case when considering our $\mathcal{L}_{\text{Rel}}$ loss, which represents a trade-off with $\mathcal{L}_{\text{ETF}}$ between the representation power of the encoder and the stability of the Simplex structure.

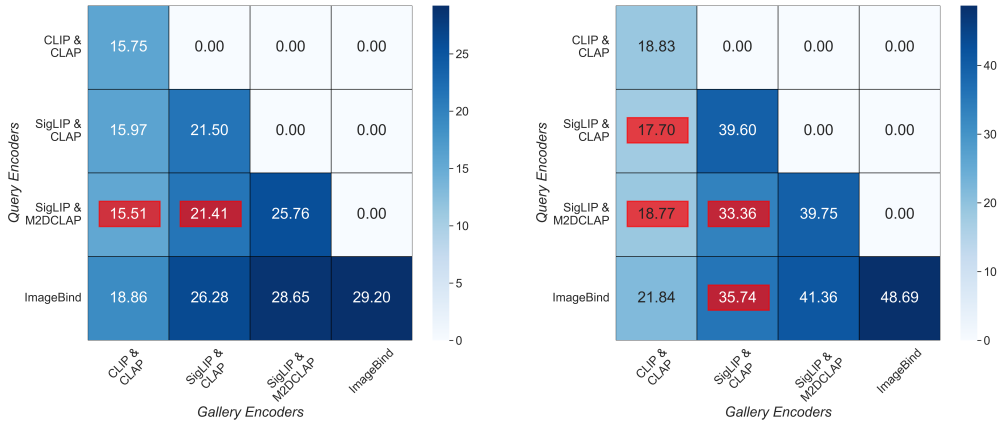
Finally, despite the stability provided by the ETF, our experiments indicate that the best reported results use replay. Removing replay substantially reduces FA, and retaining raw examples may be unsuitable where storage or data-retention constraints prohibit it. While DCAE is better than the compared baselines in the CIL setup, it is not yet fully exemplar-free. For strictly privacy-sensitive or low-resource computing, storing raw samples from previous tasks remains an impracticable solution.

H CMC MATRICES FOR AVE-CI

The non-incremental reference results on the AVE dataset are available in Table 13. We also present the zero-shot capabilities, where the backbone encoder performs the retrieval on its own, without any adapter on top. These results serve as a good approximation for the upper bounds inside the CIL context. The detailed CMC matrices for all studied methods on the AVE-CI dataset are available in Figs. 4 to 12.

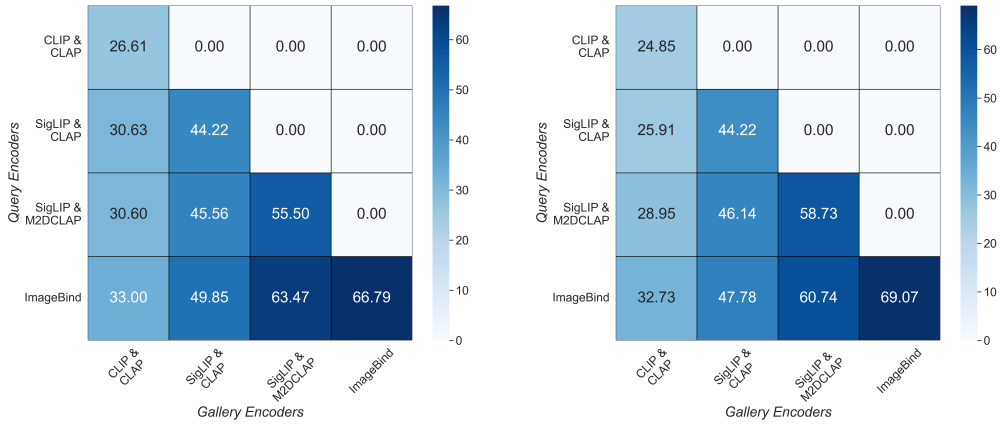
Table 13: Upper bounds for non-CIL training on AVE

Method	Retrieval	Zero-shot CMC@1	Train CMC@1
CLIP & CLAP	Audio-to-Image	5.05	34.37
	Image-to-Audio	7.66	66.42
SigLIP & M2D-CLAP	Audio-to-Image	0.94	72.17
	Image-to-Audio	2.89	78.22
ImageBind	Audio-to-Image	67.15	76.52
	Image-to-Audio	80.72	80.17



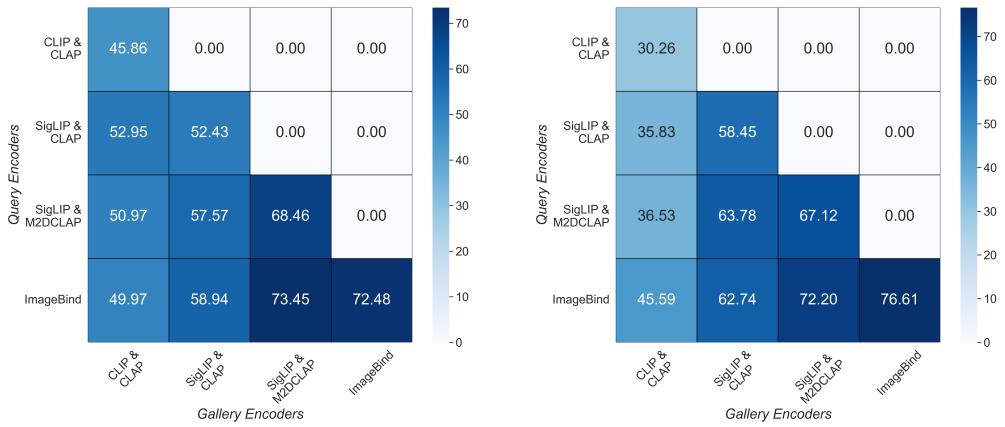
(a) Audio-to-Image (b) Image-to-Audio

Figure 4: Reference results for CLIP and CLAP (Oracle 1) on AVE-CI



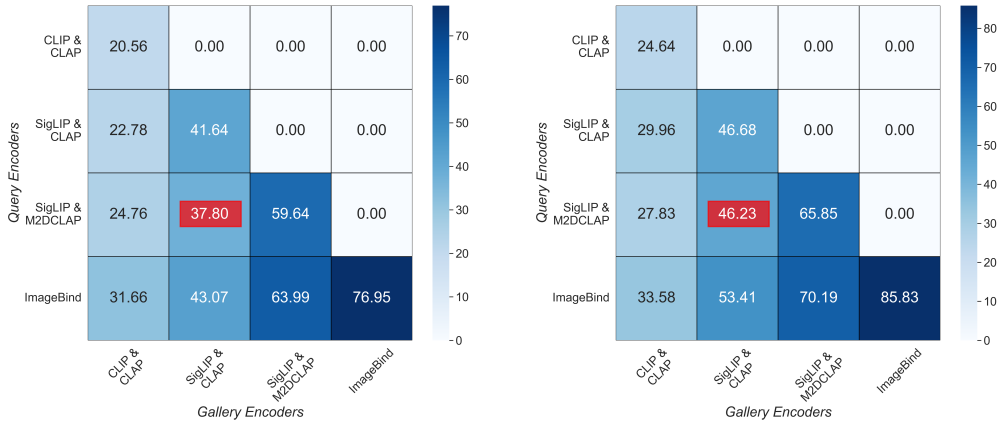
(a) Audio-to-Image (b) Image-to-Audio

Figure 5: Reference results for SigLIP and M2D-CLAP (Oracle 2) on AVE-CI



(a) Audio-to-Image (b) Image-to-Audio

Figure 6: Reference results for ImageBind (Oracle 3) on AVE-CI



(a) Audio-to-Image (b) Image-to-Audio

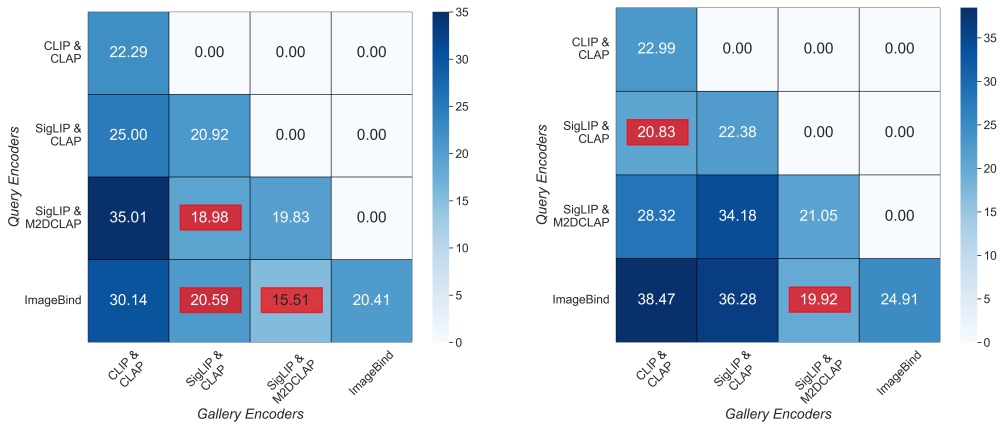
Figure 7: Result of DCAE (ours) on AVE-CI

I CMC MATRICES FOR VGGSOUND-CI

The non-incremental reference results on the VGGSound dataset are available in Table 14, following the same protocol as for the AVE dataset. The detailed CMC matrices for all studied methods on the VGGSound-CI dataset are available in Figs. 13 to 21.

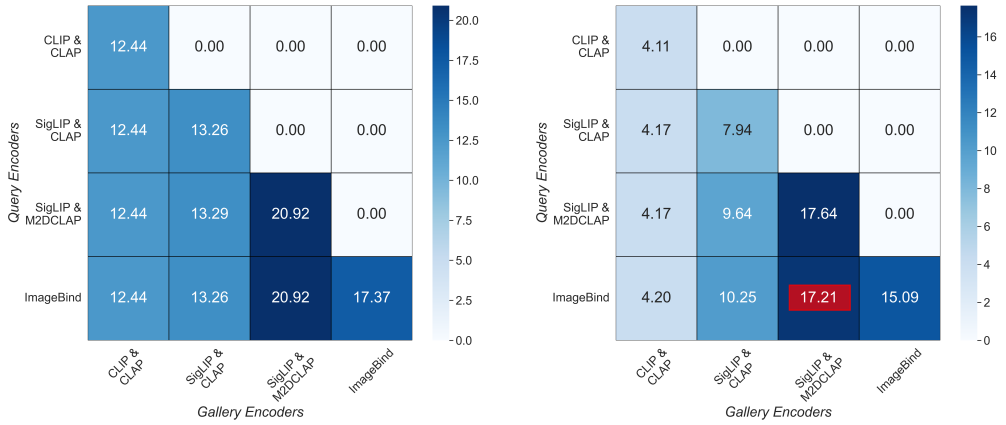
Table 14: Upper bounds for non-CIL training on VGGSound

Method	Retrieval	Zero-shot CMC@1	Train CMC@1
CLIP & CLAP	Audio-to-Image	0.84	56.49
	Image-to-Audio	0.66	53.64
SigLIP & M2D-CLAP	Audio-to-Image	1.43	69.00
	Image-to-Audio	1.05	57.84
ImageBind	Audio-to-Image	40.29	51.99
	Image-to-Audio	49.08	61.62



(a) Audio-to-Image (b) Image-to-Audio

Figure 8: Result of Gupta et al. (2026) on AVE-CI



(a) Audio-to-Image (b) Image-to-Audio

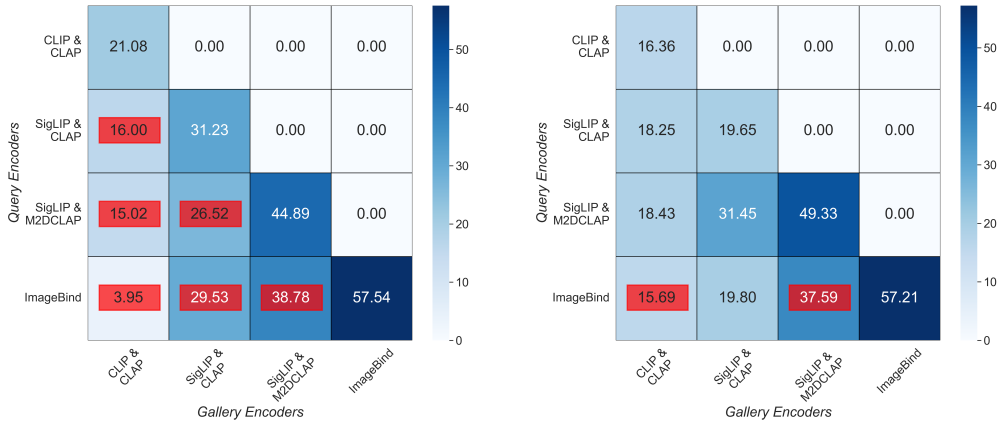
Figure 9: Result of Jang & Lim (2025) on AVE-CI

J CMC Matrices for MSCOCO-CI

The non-incremental reference results on the MSCOCO dataset are available in Table 15. The detailed CMC matrices for all studied methods on the MSCOCO-CI dataset are available in Figs. 22 to 27.

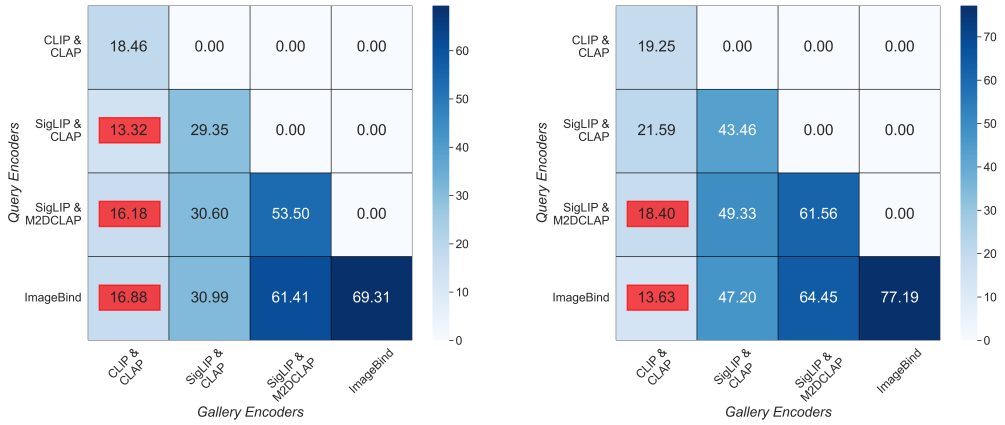
Table 15: Upper bounds for non-CIL training on MSCOCO

Method	Retrieval	Zero-shot CMC@1	Train CMC@1
CLIP & CLAP	Text-to-Image	3.93	10.50
	Image-to-Text	3.55	16.29
ImageBind	Text-to-Image	89.88	91.24
	Image-to-Text	91.00	92.57



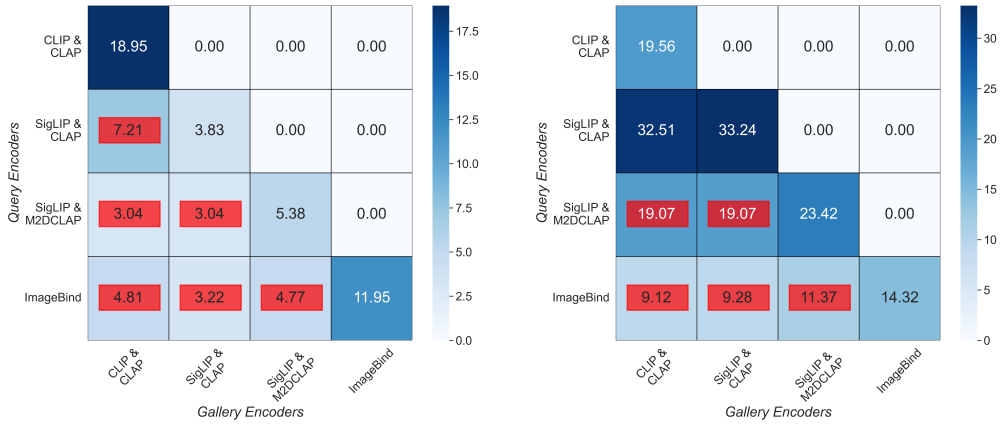
(a) Audio-to-Image (b) Image-to-Audio

Figure 10: Result of Pian et al. (2023) on AVE-CI



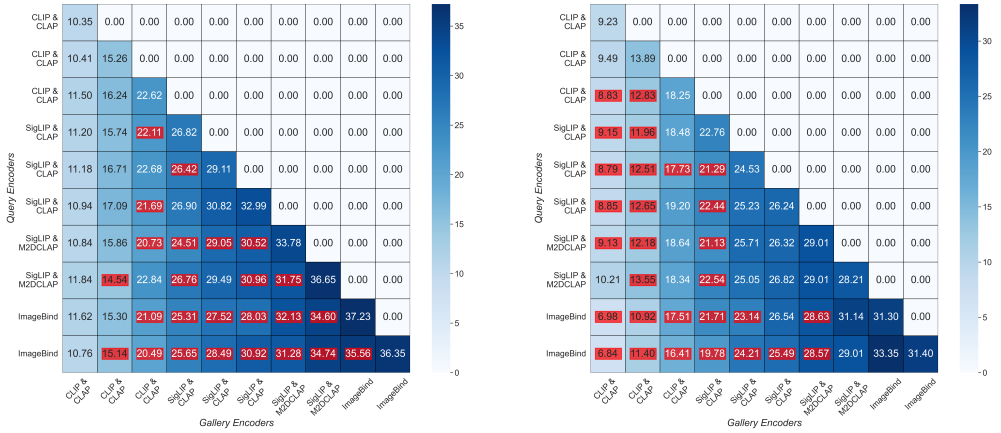
(a) Audio-to-Image (b) Image-to-Audio

Figure 11: Result of Yang et al. (2023) on AVE-CI



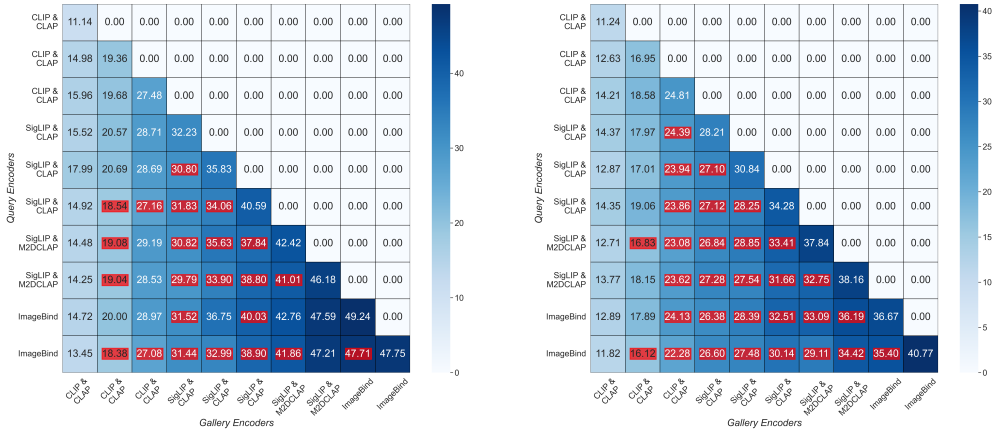
(a) Audio-to-Image (b) Image-to-Audio

Figure 12: Result of Byun et al. (2026) on AVE-CI



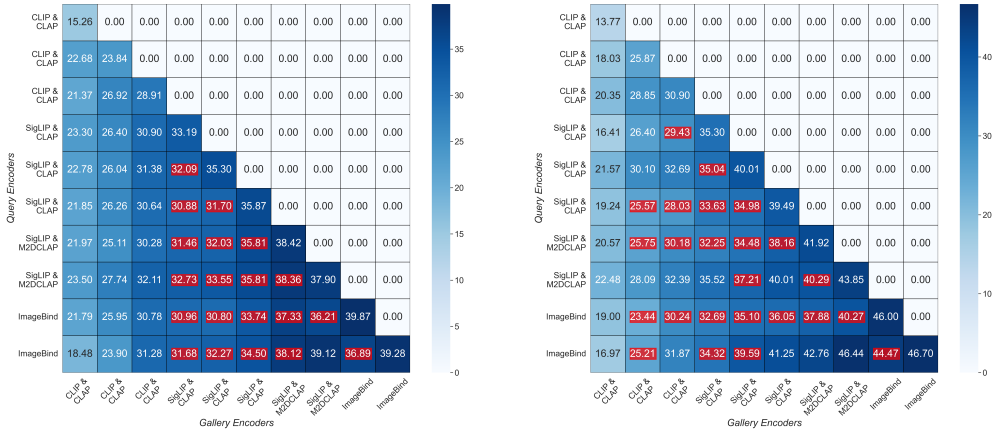
(a) Audio-to-Image (b) Image-to-Audio

Figure 13: Reference results for CLIP and CLAP (Oracle 1) on VGGSound-CI



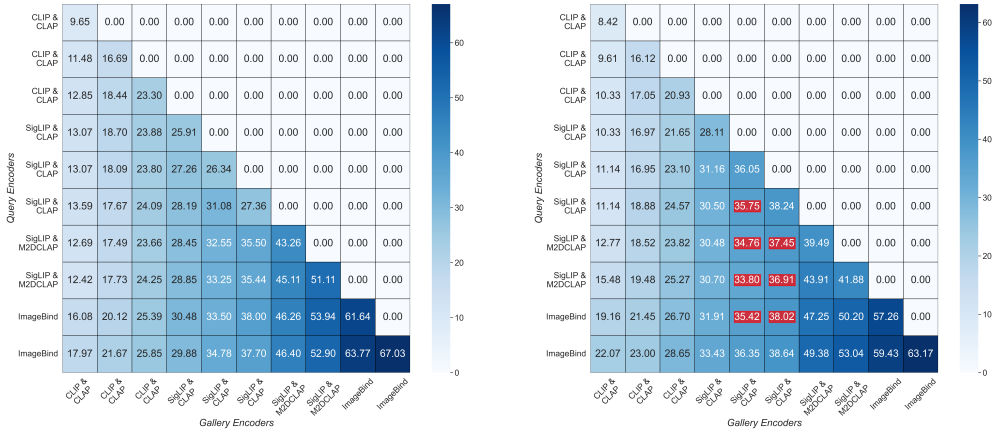
(a) Audio-to-Image (b) Image-to-Audio

Figure 14: Reference results for SigLIP and M2D-CLAP (Oracle 2) on VGGSound-CI



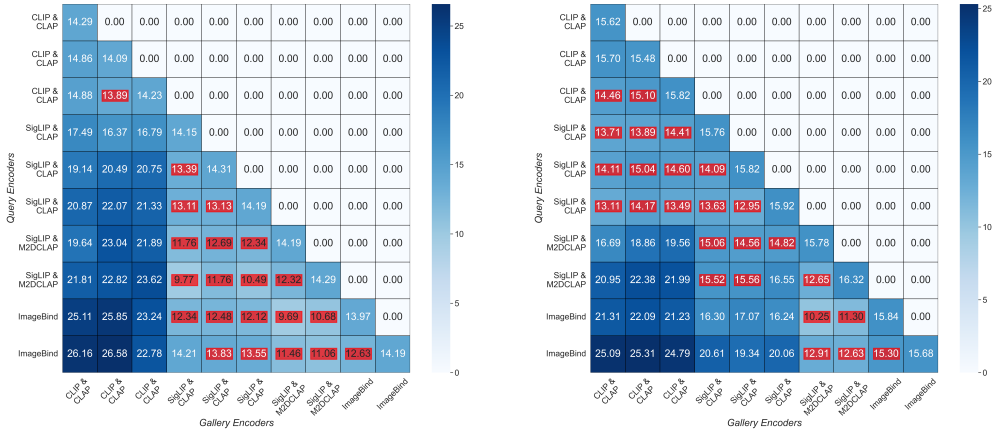
(a) Audio-to-Image (b) Image-to-Audio

Figure 15: Reference results for ImageBind (Oracle 3) on VGGSound-CI



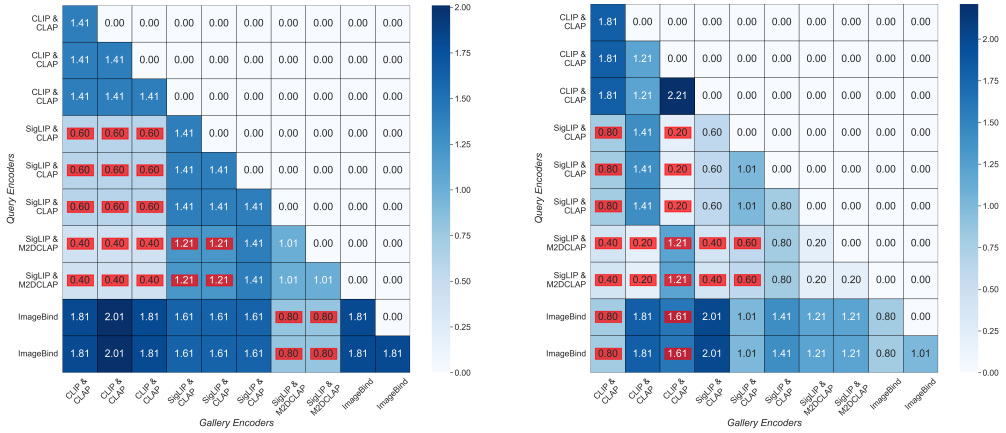
(a) Audio-to-Image (b) Image-to-Audio

Figure 16: Result of DCAE (ours) on VGGSound-CI



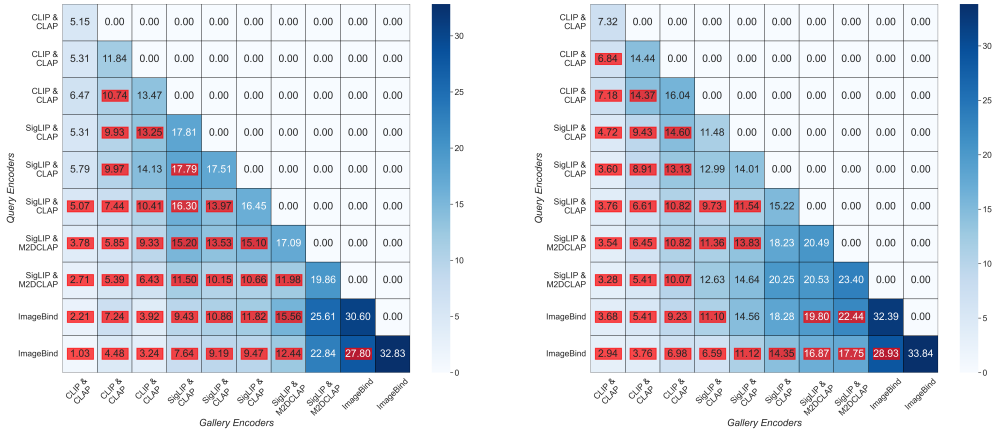
(a) Audio-to-Image (b) Image-to-Audio

Figure 17: Result of Gupta et al. (2026) on VGGSound-CI



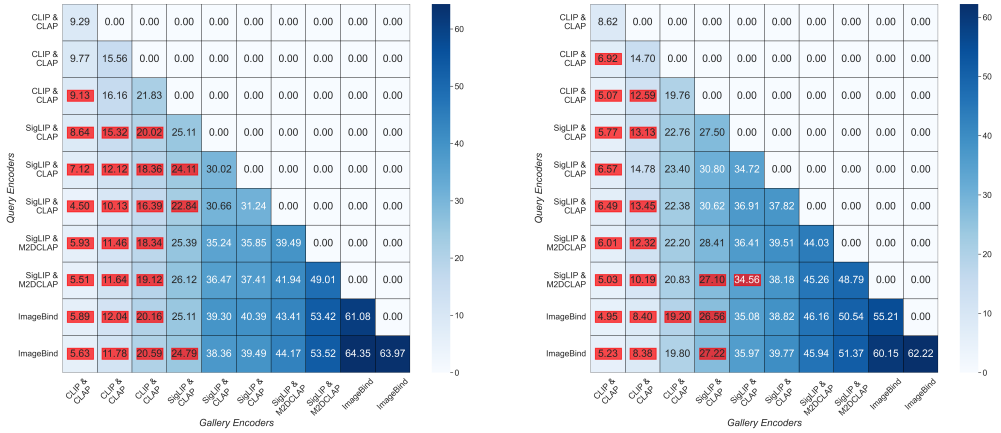
(a) Audio-to-Image (b) Image-to-Audio

Figure 18: Result of Jang & Lim (2025) on VGGSound-CI



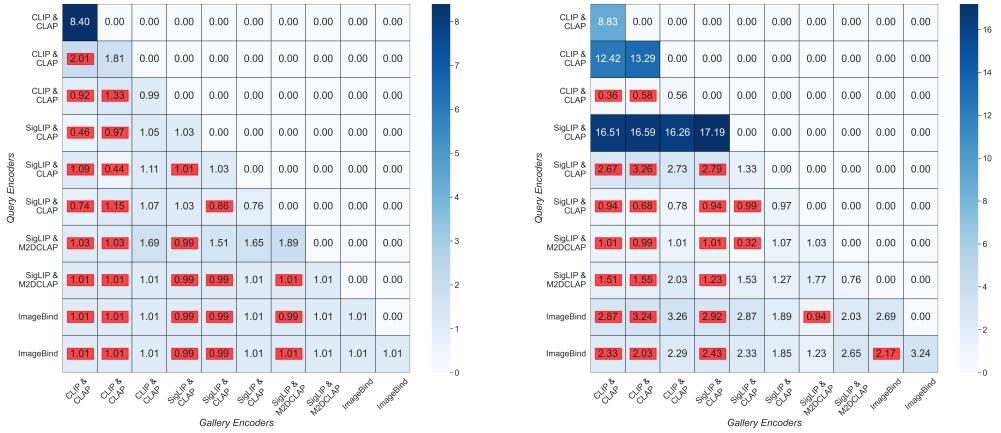
(a) Audio-to-Image (b) Image-to-Audio

Figure 19: Result of Pian et al. (2023) on VGGSound-CI



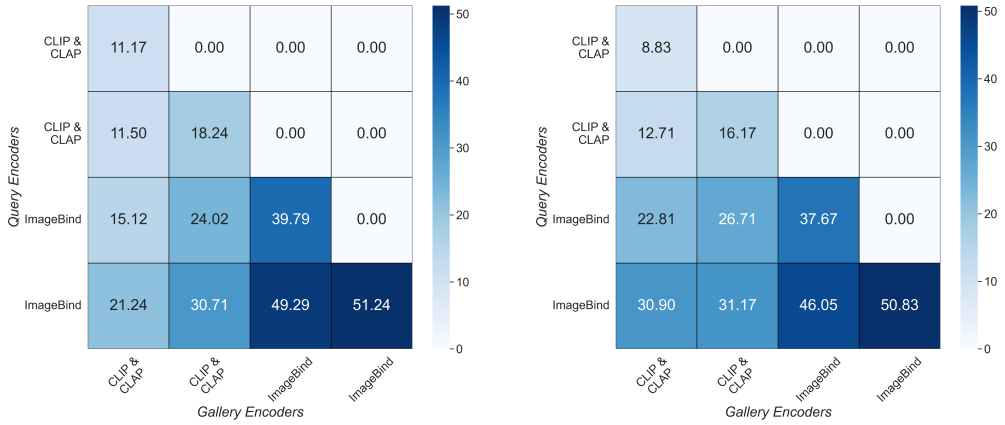
(a) Audio-to-Image (b) Image-to-Audio

Figure 20: Result of Yang et al. (2023) on VGGSound-CI



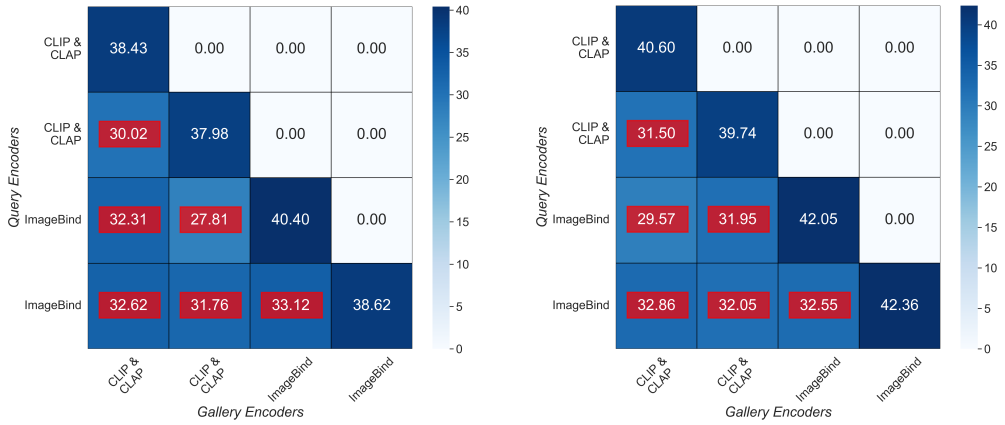
(a) Audio-to-Image (b) Image-to-Audio

Figure 21: Result of Byun et al. (2026) on VGGSound-CI



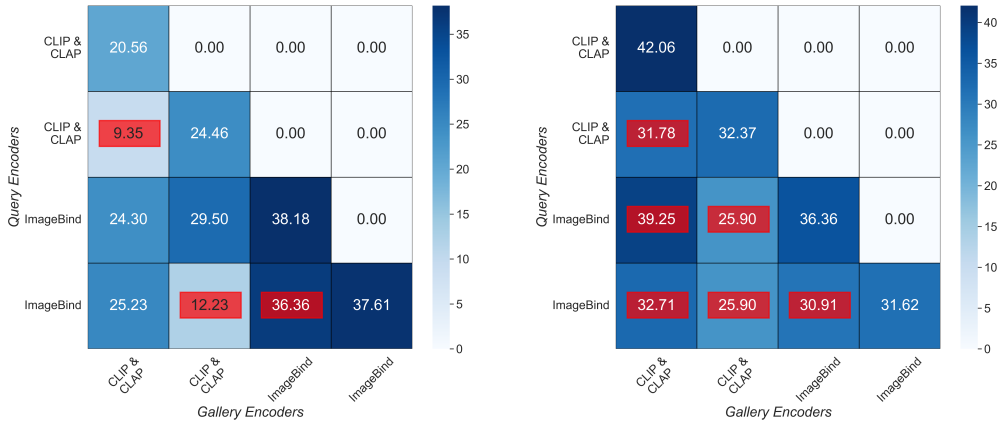
(a) Text-to-Image (b) Image-to-Text

Figure 22: Result of DCAE (ours) on MSCOCO-CI



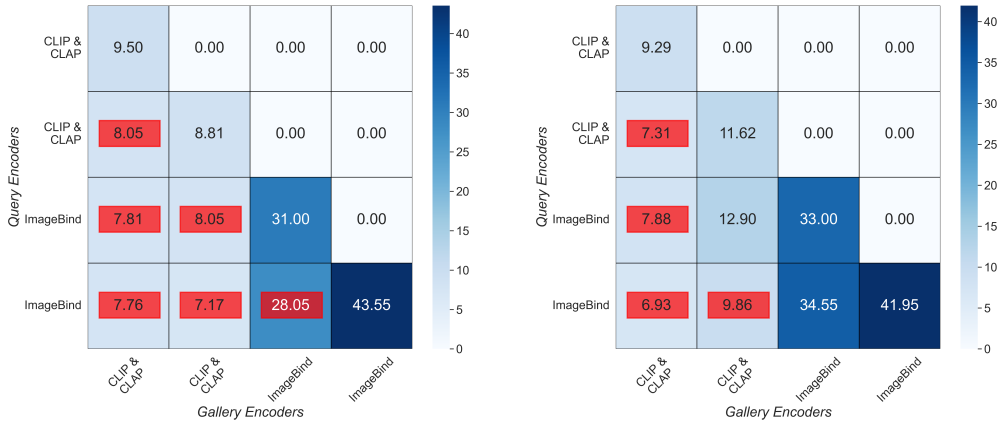
(a) Text-to-Image (b) Image-to-Text

Figure 23: Result of Gupta et al. (2026) on MSCOCO-CI



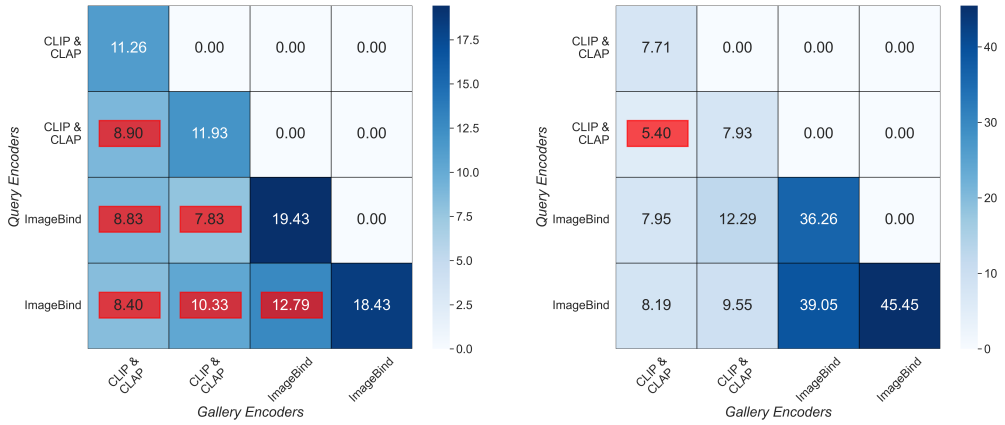
(a) Text-to-Image (b) Image-to-Text

Figure 24: Result of Jang & Lim (2025) on MSCOCO-CI



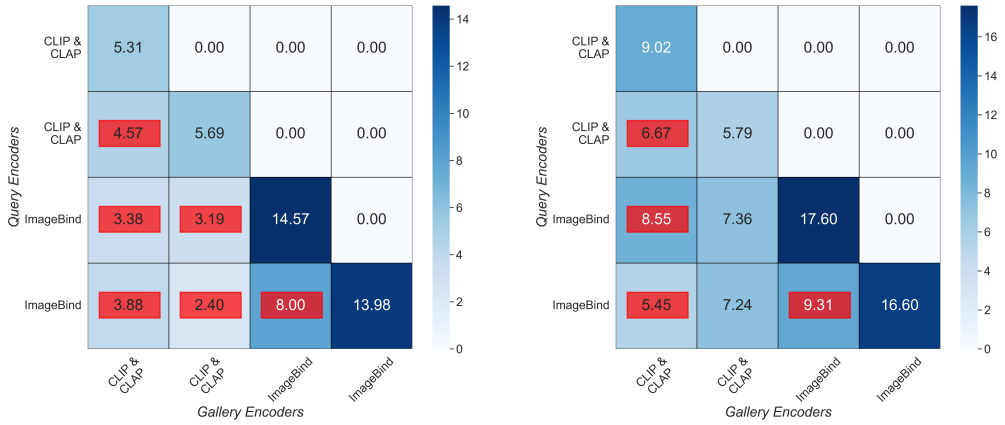
(a) Text-to-Image (b) Image-to-Text

Figure 25: Result of Pian et al. (2023) on MSCOCO-CI



(a) Text-to-Image (b) Image-to-Text

Figure 26: Result of Yang et al. (2023) on MSCOCO-CI



(a) Text-to-Image (b) Image-to-Text

Figure 27: Result of Byun et al. (2026) on MSCOCO-CI