BiCLoRA: Two-Sided Constrained LoRA for Audio-Video-Text Continual Learning

Anonymous authors
Paper under double-blind review

Abstract

Pretrained audio-video-text models align complementary auditory, visual, and linguistic information within a shared representation space, enabling unified recognition of real-world events. In deployment, these models must continually acquire concepts beyond their pretraining data, often without retaining past observations because of privacy and storage constraints. We formulate this problem as exemplar-free Audio-Video-Text Continual Learning (AVT-CL), which extends an aligned multimodal space while preserving prior knowledge and cross-modal alignment. We identify two failure modes of continual low-rank adaptation: *historical interference*, where the input factor overlaps with activation subspaces occupied by previous sessions, and *alignment drift*, where a freely trained output factor writes new information into arbitrary shared-space directions, even when the input side is protected. To address both, we propose *Bilaterally Constrained Low-Rank Adaptation* (**BiCLoRA**), which fixes both factors before optimization. A history-safe input basis derived from a bounded eigensummary of past projection inputs restricts where updates may act, while a frozen output basis spanning the leading directions of current-session representations restricts where new information may be written. Only a small coupling kernel for each modality is optimized and subsequently merged into the current projection. Across three pretrained AVT backbones and three benchmarks, BiCLoRA improves macro final average accuracy from 69.08 to 72.21 and backward transfer from -13.17 to -7.72 over the strongest matched adaptation baseline. Ablations confirm that the input constraint protects historical representations, the output constraint preserves cross-modal alignment, and effective retention requires both.

1 Introduction

Real-world events are inherently multimodal: their identities and dynamics are jointly conveyed by sound, visual appearance, and language, with each modality providing information unavailable to the others. Recent pretrained audio-video-text (AVT) models, including ImageBind, LanguageBind, and EBind, align these complementary signals within a shared representation space, enabling unified recognition across modalities (Girdhar et al., 2023; Zhu et al., 2024; Broadbent et al., 2025). In deployment, however, these models must continually acquire concepts and adapt to environments beyond their pretraining data, often without retaining past observations because of privacy and storage constraints. This motivates exemplar-free Audio-Video-Text Continual Learning (AVT-CL), which extends an aligned multimodal space while preserving previously acquired knowledge and cross-modal alignment. We consider a practical protocol in which each session provides paired audio-video observations and class names for newly arriving concepts, after

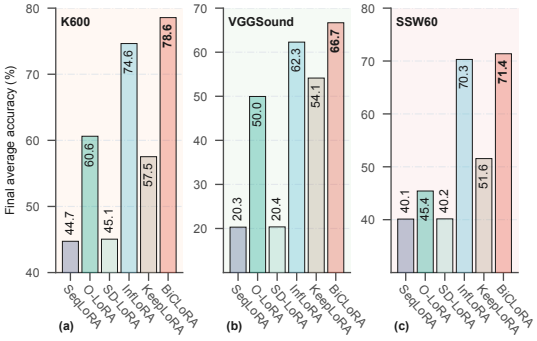


Figure 1: **Low-rank CL on an aligned tri-modal space.** Final average accuracy per benchmark, mean over three backbones (two for KeepLoRA).

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.

Table 1: **Comparison of representative continual learning settings.** AVT-CL learns from audio, video, and text without exemplar replay, while inference relies only on audio and video.

Setting	Training modalities	Inference modalities	Replay	Role of text
Visual CL (e.g., Liang & Li, 2024)	image	image	no	absent
Vision-Language CL (e.g., Jha et al., 2024)	image, text	image, text	no	semantic classifier
Audio-Visual CL (e.g., Pian et al., 2023)	audio, video	audio, video	yes	absent
Audio-Video-Text CL (Ours)	audio, video, text	audio, video	no	training-time guidance

which its observations and instance-level features are discarded. Text provides semantic guidance during training, while inference requires recognizing all concepts encountered so far from audio and video alone. Table 1 contrasts this protocol with existing CL settings.

Parameter-efficient tuning has become a common approach to CL with pretrained models, enabling adaptation through a small set of trainable parameters while keeping most of the backbone frozen. Recent subspace-based methods further mitigate interference by constraining update directions or protecting historical representations (Saha et al., 2021; Liang & Li, 2024; Luo et al., 2026). In a low-rank update, these mechanisms primarily constrain the *input* factor, which determines which activations the update responds to. The *output* factor, which determines where the resulting changes are written in representation space, is typically trained without an analogous constraint. This line of research has focused mainly on vision and vision-language models with a single perceptual branch, where text, when present, usually serves as a classifier or semantic anchor. Audio-visual CL instead considers the joint evolution of auditory and visual representations but generally excludes text from continual adaptation (Pian et al., 2023; Mo et al., 2023; Yue et al., 2024).[1] These works therefore leave a central question: how can a pretrained AVT space be continually adapted without overwriting historical representations or disrupting the cross-modal geometry that binds its modalities?

To characterize this gap, we conduct a controlled empirical analysis and identify one failure mode on each side of a low-rank update. The input factor determines which activation directions an update responds to, while the output factor determines where the resulting changes are written in the shared representation space. On the input side, *historical interference* arises when the update directions overlap with activation subspaces occupied by previous sessions. Sequential LoRA (SeqLoRA) consequently perturbs historical activations with up to 78 times the energy of a norm-matched isotropic update and suffers severe forgetting (Table 2; Fig. 1). On the output side, *alignment drift* arises when a freely trained output factor redirects changes across shared-space directions, disrupting the geometry that aligns audio, video, and text. Input-side protection alone does not prevent this drift: in the fine-grained setting, the audio-video similarity of first-session samples becomes negative even when the input basis is orthogonal to the retained historical subspace. Effective AVT-CL must therefore constrain both where an update may act and where the resulting changes may lie.

In this regard, we propose *Bilaterally Constrained Low-Rank Adaptation* (**BiCLoRA**), which fixes data-dependent bases on both sides of each low-rank update before optimization. To mitigate historical interference, BiCLoRA constructs a *history-safe input basis* by projecting current activations onto the orthogonal complement of a bounded eigensummary of historical projection inputs. This restricts updates to directions outside the retained historical subspace while keeping persistent state bounded. To mitigate alignment drift, a *frozen output basis* spans the leading directions of the current-session representations produced by the consolidated model, confining changes to an output subspace already supported before optimization. With both bases fixed, only a small coupling kernel between them is optimized for each modality. The learned updates are then merged into the current projections, after which all temporary adaptation components are discarded.

We evaluate BiCLoRA with ImageBind, LanguageBind, and EBind on Kinetics-600, VGGSound, and SSW60. Across the nine backbone-benchmark settings, BiCLoRA achieves the best backward transfer in all and the highest final average accuracy in eight. Over the strongest matched adaptation baseline, it improves macro final average accuracy by 3.13 points, from 69.08 to 72.21, and backward transfer by 5.45 points, from -13.17 to -7.72, using only 128 trainable adaptation parameters per session. Ablations show that the history-safe input basis protects historical representations, the frozen output basis preserves cross-modal alignment, and effective retention requires both.

[1] Due to the page limit, a comprehensive discussion of related work is provided in Appendix A.

Our contributions are threefold. (1) We formulate exemplar-free AVT-CL and identify two factor-specific failure modes of continual low-rank adaptation: input-side historical interference and output-side alignment drift that persists under input-side protection. (2) We propose BiCLoRA, which combines a bounded-state, history-safe input basis with a frozen output basis and optimizes only a small coupling kernel, with a bound on the resulting perturbation of historical activations. (3) Experiments across three AVT backbones and three benchmarks demonstrate improved accuracy and retention over matched adaptation baselines and validate the distinct roles of both constraints.

2 Audio-Video-Text Continual Learning

Pretrained audio-video-text models align auditory, visual, and linguistic semantics within a shared representation space. Audio-Video-Text Continual Learning (AVT-CL) studies how this space can be continually extended as new concepts arrive while retaining previously acquired knowledge and cross-modal alignment without storing past observations. We formalize the exemplar-free AVT-CL protocol below and present an in-depth empirical analysis.

2.1 Problem Formulation

Learning proceeds over a sequence of T sessions. At session t, the learner receives a dataset of paired audio-video observations, $\mathcal{D}_t = \{(x_i^a, x_i^v, y_i)\}_{i=1}^{N_t}$, where N_t is the number of training samples and $y_i \in \mathcal{C}_t$ is the label of sample i within the current class set $\mathcal{C}_t$. The learner is also given the name of each class $c \in \mathcal{C}_t$, which is converted into a textual input x_c^{txt} using a fixed prompt template. Class sets are disjoint across sessions, i.e., $\mathcal{C}_t \cap \mathcal{C}_s = \varnothing$ for all $t \neq s$. Once a session ends, its raw observations and instance-level features are discarded. The pretrained model maps all three modalities into a common p-dimensional representation space. For each perceptual modality $m \in \{a, v\}$, where a and v denote audio and video, respectively, a frozen encoder trunk maps x_i^m to a projection input $h_i^m \in \mathbb{R}^{d_m}$. An adaptable linear projection $W_{m,t} \in \mathbb{R}^{p \times d_m}$ then maps this activation into the shared space. Let $W_{m,0}$ denote the pretrained projection. During session t, the learner estimates a modality-specific update $\Delta W_{m,t} \in \mathbb{R}^{p \times d_m}$ and consolidates it as

$$W_{m,t} = W_{m,t-1} + \Delta W_{m,t}. \tag{1}$$

All encoder-trunk parameters remain frozen throughout the continual stream, so representation adaptation occurs only through $\Delta W_{a,t}$ and $\Delta W_{v,t}$. The text encoder f_{txt} also remains frozen.

After session t, the modality-specific and joint audio-video representations are

$$z_{i,t}^m = \text{Norm}(W_{m,t} h_i^m), \qquad j_{i,t} = \text{Norm}\big(\tfrac{1}{2}\left(z_{i,t}^a + z_{i,t}^v\right)\big), \tag{2}$$

where $\text{Norm}(u) = u/\|u\|_2$ denotes ℓ_2 normalization. The representation of class c is

$$t_c = \text{Norm}\big(f_{\text{txt}}\big(x_c^{\text{txt}}\big)\big), \tag{3}$$

and is used only as semantic guidance during training. When the session is clear from context, we omit its index from $z_{i,t}^m$ and $j_{i,t}$. In the method section, W_m denotes the consolidated projection $W_{m,t-1}$ entering session t, and ΔW_m denotes the corresponding update $\Delta W_{m,t}$.

Following common practice in CL with pretrained models, the joint representation $j_{i,t}$ is passed to an analytic classifier that is updated across sessions using aggregate sufficient statistics (McDonnell et al., 2023). All methods in Table 2 share the same classifier design, update rule, and hyperparameters, allowing performance differences to be attributed to representation adaptation. After session t, the model must recognize every class in $\mathcal{C}_{1:t} = \bigcup_{s=1}^{t} \mathcal{C}_s$ using audio and video alone. Exemplar replay, task identity, test-time text, and stored historical instance-level features are not permitted, although model parameters and aggregate non-instance-level statistics may persist across sessions. The central learning problem is therefore to estimate $\Delta W_{a,t}$ and $\Delta W_{v,t}$ from the current session while retaining historical representations and the shared geometry among audio, video, and text. Setting $\Delta W_{m,t} = 0$ recovers the frozen pretrained model, which serves as the reference for representation drift.

2.2 Empirical Analysis of AVT-CL

To understand why continual low-rank adaptation can degrade an aligned AVT space, we examine the two factors of a rank-r update for modality m, $\Delta W_m = B_m A_m^\top$, where $A_m \in \mathbb{R}^{d_m \times r}$ is the

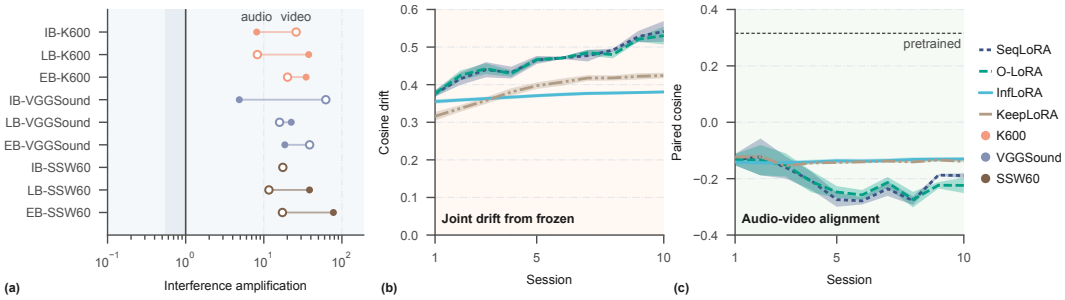


Figure 2: **Two failure modes of continual low-rank adaptation. (a)** Perturbation of past projection inputs by SeqLoRA relative to a norm-matched isotropic update (filled: audio, open: video). **(b)** Drift of first-session joint representations from the pretrained model (ImageBind–SSW60) and **(c)** their audio-video alignment (dashed: pretrained). Mean ± s.d. over three seeds.

input factor and $B_m \in \mathbb{R}^{p\times r}$ is the output factor. The input factor determines which activation directions the update responds to, whereas the output factor determines which shared-space directions receive the resulting changes. We identify a distinct failure mode on each side (Fig. 2).

Challenge 1: historical interference on the input side. For a historical projection input h^m, the induced change is $\Delta W_m h^m = B_m(A_m^\top h^m)$. Low rank limits the dimensionality of the update but does not prevent its input directions from overlapping with activation subspaces used by previous sessions. To quantify this overlap, let $W_{m,t}$ denote the projection of modality m after session t and $\Delta W_{m,t} = W_{m,t} - W_{m,t-1}$ its session-wise update. We summarize the projection inputs from sessions 1 through $t-1$ by their dominant eigenspace and use two scale-independent measures: the fraction of update energy aligned with the top-16 historical directions and the perturbation energy on historical inputs relative to a norm-matched isotropic random update. Across the nine backbone-benchmark settings, SeqLoRA amplifies historical perturbations by factors of 4.9 to 77.8 (Fig. 2a) and concentrates up to 58% of its update energy in dominant historical directions. By the end of the stream, its first-session audio and video representations drift by 0.22 to 0.64 in cosine distance from the frozen model. Together with Fig. 1, these results show that low rank alone does not protect historical representations, and the retained historical input subspace must be explicitly protected.

Challenge 2: alignment drift through the output side. Input-side protection eliminates overlap with the retained historical subspace but leaves the output factor freely trainable. Any residual signal passing through the input filter can therefore be redirected by B_m across the shared output space, whose directions jointly encode previously acquired concepts and the correspondence among audio, video, and text. On ImageBind–SSW60, with input-side protection alone, as represented by InfLoRA and KeepLoRA, the cosine drift of first-session joint representations remains 0.38 and 0.42, compared with 0.54 under SeqLoRA (Fig. 2b). The audio-video similarity of the same samples also falls from 0.32 in the frozen model to -0.13 and -0.14, remaining negative as under the unprotected update (-0.19; Fig. 2c). Holding the input constraint fixed and additionally freezing the output basis restores audio-video similarity to 0.25 and reduces drift to 0.02. Even without input-side protection, freezing the output basis keeps the similarity positive at 0.17 and reduces drift to 0.28 (Fig. 4). These comparisons isolate the freely trained output factor as a separate source of alignment drift. Effective AVT-CL therefore requires complementary constraints on both sides: the input basis must avoid retained historical directions, and the output basis must be fixed before optimization rather than trained freely.

3 BiCLoRA: Bilaterally Constrained Low-Rank Adaptation

Our analysis motivates *Bilaterally Constrained Low-Rank Adaptation* (**BiCLoRA**), which addresses the two challenges through complementary constraints on both sides of each low-rank update. To prevent historical interference, BiCLoRA constructs a history-safe input basis orthogonal to the subspace captured by a bounded eigensummary of previous projection inputs. To prevent alignment drift, it constructs a frozen output basis from the leading directions of the current-session representations before optimization. The input constraint determines where an update may act, and the output constraint where the resulting information may be written.

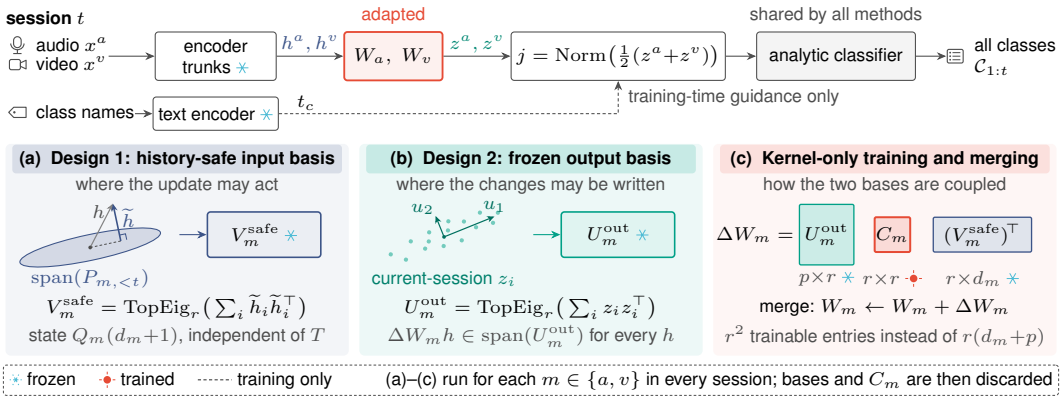


Figure 3: **Overview of BiCLoRA.** Frozen encoders map audio and video into a shared space, and only their projections W_a and W_v are adapted. Text guides training but is not used at inference. **(a)** A bounded eigensummary of historical inputs defines a history-safe input basis in its orthogonal complement. **(b)** The leading directions of the current-session representations define a frozen output basis that confines changes to a fixed r-dimensional subspace. **(c)** Only the small coupling kernel between the two bases is optimized, after which the update is merged into the projection.

For each modality $m \in \{a, v\}$, BiCLoRA parameterizes the projection update as

$$\Delta W_m = U_m^{\mathrm{out}} C_m \left(V_m^{\mathrm{safe}}\right)^\top, \tag{4}$$

where $V_m^{\mathrm{safe}} \in \mathbb{R}^{d_m \times r}$ is the history-safe input basis, $U_m^{\mathrm{out}} \in \mathbb{R}^{p \times r}$ is the frozen output basis, and r is the adaptation rank. Both bases are constructed once at the beginning of each session and remain fixed during optimization. Only the coupling kernel $C_m \in \mathbb{R}^{r \times r}$ is learned, determining how the selected input directions map to the constrained output subspace without changing either basis. After optimization, the resulting update is merged into the current projection. Fig. 3 provides an overview of the complete framework.

3.1 Design 1: History-Safe Input Basis

The input-side constraint keeps new updates off activation directions retained from previous sessions. For each modality, BiCLoRA maintains a bounded eigensummary of historical projection inputs and constructs the current input basis from directions orthogonal to the retained history.

Bounded history summary. Let Q_m denote the history-rank budget for modality m. Before session t, BiCLoRA retains a recursively truncated eigensummary $(P_{m,<t}, \Lambda_{m,<t})$ of the uncentered second moment of previous projection inputs. The columns of $P_{m,<t}$ contain at most Q_m orthonormal eigenvectors, and the diagonal entries of $\Lambda_{m,<t}$ are their corresponding eigenvalues. After session t, its projection inputs are incorporated and the summary is truncated back to the prescribed rank:

$$(P_{m,\le t}, \Lambda_{m,\le t}) = \mathrm{TruncEig}_{Q_m}\left(P_{m,<t}\Lambda_{m,<t}P_{m,<t}^\top + \sum_{i=1}^{N_t} h_i^m (h_i^m)^\top\right), \tag{5}$$

where $\mathrm{TruncEig}_{Q_m}$ returns the leading Q_m eigenpairs, and $(P_{m,\le t}, \Lambda_{m,\le t})$ denotes the updated summary containing information through session t. The summary is initialized as empty before the first session. Its persistent storage is at most $Q_m(d_m + 1)$ scalars per modality and therefore does not grow with the number of sessions. Because the encoder trunks are frozen, projection inputs from different sessions remain in the same coordinate system.

Constructing the history-safe basis. At the beginning of session t, BiCLoRA removes from each current projection input the component captured by the retained history and extracts the leading directions of the residuals:

$$\tilde{h}_i^m = \left(I_{d_m} - P_{m,<t}P_{m,<t}^\top\right) h_i^m, \qquad V_m^{\mathrm{safe}} = \mathrm{TopEig}_r\left(\sum_{i=1}^{N_t} \tilde{h}_i^m (\tilde{h}_i^m)^\top\right), \tag{6}$$

where I_{d_m} is the d_m-dimensional identity matrix, $\widetilde{h}_i^m$ is the component of h_i^m outside the retained historical subspace, and TopEig_r returns the leading r orthonormal eigenvectors. For numerical stability, V_m^{safe} is reprojected and orthonormalized to ensure that $P_{m,<t}^\top V_m^{\mathrm{safe}} = 0$.

This orthogonality directly limits interference with previous inputs. For a historical projection input h_{old}^m, write $h_{\mathrm{old}}^m = P_{m,<t}\alpha_m + \epsilon_m$, where $\alpha_m = P_{m,<t}^\top h_{\mathrm{old}}^m$ contains its coordinates in the retained subspace and $\epsilon_m = (I_{d_m} - P_{m,<t}P_{m,<t}^\top)h_{\mathrm{old}}^m$ is the residual not captured by the bounded summary. Using the update in Eq. 4 and the orthonormality of both bases gives

$$\|\Delta W_m h_{\mathrm{old}}^m\|_2 \le \|C_m\|_2 \, \|\epsilon_m\|_2 \,, \tag{7}$$

where $\|C_m\|_2$ denotes the spectral norm of C_m. This bound applies to the projection output before feature normalization. Historical inputs contained entirely in the retained subspace are unaffected; otherwise, the perturbation scales with the omitted residual and the spectral norm of the kernel. This input-side design determines *where* an update may act. The output-side design next constrains where the resulting changes may lie.

3.2 Design 2: Frozen Output Basis and Kernel-Only Training

The history-safe input basis determines which activation directions may trigger an update but does not constrain where the resulting changes are written in the shared space. A freely trained output factor can still redirect them across output directions and thereby disrupt cross-modal alignment. BiCLoRA removes this remaining degree of freedom by fixing the output basis before optimization.

For each modality $m \in \{a, v\}$, BiCLoRA constructs the output basis from the leading directions of the current-session representations produced by the consolidated model:

$$U_m^{\mathrm{out}} = \mathrm{TopEig}_r \left(\sum_{i=1}^{N_t} z_i^m (z_i^m)^\top \right), \tag{8}$$

where $z_i^m = \mathrm{Norm}(W_m h_i^m)$ is computed using the consolidated projection W_m before optimization. The resulting basis U_m^{out} remains fixed throughout the session. Consequently, for any input h^m, the pre-normalization change $\Delta W_m h^m$ lies in $\mathrm{span}(U_m^{\mathrm{out}})$, an output subspace already supported by the current-session representations. The coupling kernel C_m learns how coordinates in the history-safe input basis are mapped to this fixed output basis but cannot change the subspace.

The history-safe input basis and frozen output basis leave only C_m trainable among the parameters defining ΔW_m. This reduces the number of trainable adaptation parameters from $r(d_m + p)$ for a conventional LoRA factor pair to r^2 per modality. We use $r = 8$ by default, yielding $2r^2 = 128$ trainable adaptation parameters across the audio and video branches. Thus, the parameter reduction follows directly from the two-sided constraint: V_m^{safe} restricts which input directions the update may respond to, while U_m^{out} fixes where the resulting changes may lie in the shared output space.

3.3 Optimization and Consolidation

Each coupling kernel is initialized as $C_m = 0$, and hence $\Delta W_m = 0$, so every session starts exactly from the consolidated model. Let j_i^{adp} denote the joint audio-video representation defined in Eq. 2 after replacing each projection W_m with its adapted counterpart $W_m + \Delta W_m$. We introduce a temporary bias-free linear classifier $g_t : \mathbb{R}^p \to \mathbb{R}^{|\mathcal{C}_t|}$ over the current class set $\mathcal{C}_t$, with its class weights initialized from the corresponding class-text representations. The coupling kernels and temporary classifier are optimized using

$$\mathcal{L} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left[\mathrm{CE}\big(g_t(j_i^{\mathrm{adp}}), y_i\big) + \lambda_{\mathrm{text}} \, \mathrm{CE}\left(\left[(j_i^{\mathrm{adp}})^\top t_c / \tau \right]_{c \in \mathcal{C}_t}, y_i \right) \right], \tag{9}$$

where CE denotes cross-entropy loss, λ_{text} controls the strength of text-based supervision, and τ is the temperature applied to the class-text logits. The first term learns to discriminate the current classes, while the second uses the frozen class-text representations $\{t_c\}_{c \in \mathcal{C}_t}$ to preserve semantic compatibility with the text space. Only C_a, C_v, and the temporary classifier g_t receive gradients. The encoder trunks, modality projections, text encoder, and both bases remain fixed during optimization. Because the admissible input and output directions are already restricted by construction, no additional subspace-preservation penalty is required.

After optimization, each update is consolidated into its projection as $W_m \leftarrow W_m + \Delta W_m$. The final joint representations of the current session are used to update the common analytic classifier, while the projection inputs $\{h_i^m\}_{i=1}^{N_t}$ are incorporated into the bounded history summary through Eq. 5. The input and output bases, coupling kernels, and temporary classifier are then discarded. At inference, the model uses only the frozen encoder trunks, consolidated projections, and common analytic classifier. The text encoder and temporary adaptation components are not required.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate on three audio-video-text benchmarks covering human action, sound source, and fine-grained species recognition. For Kinetics-600 (Carreira et al., 2018), we use the first 200 classes in 20 ten-class sessions, retain up to 50 training clips per class, and evaluate on 9,579 validation clips after removing 333 clips that also appear in the training subset. For VGG-Sound (Chen et al., 2020), we use 200 classes in 40 five-class sessions, retain up to 100 unique training clips per class, and evaluate on 4,000 test clips. For SSW60 (Van Horn et al., 2022), we use all 60 bird species in ten six-class sessions, follow the official video split with audio extracted from the corresponding videos, and evaluate on 1,938 test clips.

Backbones. We consider three pretrained audio-video-text backbones. For ImageBind (Girdhar et al., 2023), we adapt the final audio and visual projections of sizes 1024×768 and 1024×1280, respectively. For LanguageBind (Zhu et al., 2024), we adapt the corresponding 768×1024 projections in both branches. EBind (Broadbent et al., 2025) contains a two-layer MLP audio projector but outputs video representations directly from its visual encoder. We keep the audio MLP frozen and adapt a zero-initialized linear residual on its input, so that the audio representation is $\mathrm{Norm}(\mathrm{MLP}(h_i^a) + W_{a,t} h_i^a)$ and the update $\Delta W_{a,t}$ still acts linearly on h_i^a; at the visual interface we add an identity-initialized linear projection. All other backbone parameters remain frozen.

Baselines. We compare BiCLoRA with low-rank CL methods under a common classifier interface: SeqLoRA (Hu et al., 2021), O-LoRA (Wang et al., 2023), SD-LoRA (Wu et al., 2025), InfLoRA (Liang & Li, 2024), and KeepLoRA (Luo et al., 2026). All are applied to the same audio and video projection layers as BiCLoRA. Vision-language CL methods are examined separately with their own prediction heads in Appendix B.

Implementation details. For BiCLoRA, we use rank $r = 8$ per modality ($R = 16$ in total, matching the LoRA baselines) and history rank $Q_m = 64$. We set $\tau = 0.2$ and $\lambda_{\text{text}} = 0.05$. All methods in Table 2 are trained for one epoch using AdamW with a learning rate of 5×10^{-4}. Across all backbone-benchmark settings, the common analytic classifier uses a 2,048-dimensional random projection and ridge coefficient 10^3. Regarding evaluation metrics, let $A_{k,t}$ denote the accuracy on the classes introduced in session k after training through session t. We report the final average accuracy $\mathrm{FAA} = \frac{1}{T}\sum_{k=1}^{T} A_{k,T}$ and backward transfer $\mathrm{BWT} = \frac{1}{T-1}\sum_{k=1}^{T-1}(A_{k,T} - A_{k,k})$.

4.2 EXPERIMENTAL RESULTS

Comparison of overall performance. Table 2 compares all adaptation methods under the common analytic classifier. SeqLoRA and SD-LoRA suffer severe forgetting, especially on the 40-session VGGSound stream, where their FAA ranges from 17.91 to 24.51, compared with 63.39–69.02 for BiCLoRA. O-LoRA and KeepLoRA, where applicable, improve retention but consistently trail InfLoRA, the strongest low-rank baseline. BiCLoRA achieves the best BWT in all nine backbone-benchmark settings and the highest FAA in eight. Overall, it improves macro FAA over InfLoRA by 3.13 points, from 69.08 to 72.21, and BWT by 5.45 points, from -13.17 to -7.72. The performance gains are consistent across datasets and backbones. Averaged over the three backbones, BiCLoRA improves FAA/BWT by $3.92/5.49$ points on Kinetics-600, $4.39/8.62$ on VGGSound, and $1.08/2.24$ on SSW60. The larger gains on the longer Kinetics-600 and VGGSound streams are consistent with accumulated historical interference. BiCLoRA achieves these results with only 128 trainable adaptation parameters per session, compared with 16,384 for InfLoRA and KeepLoRA, while maintaining bounded state and no inference cost beyond the merged projections (Table 4).

Table 2: **Comparison of overall performance across backbones and benchmarks.** Final average accuracy (FAA) and backward transfer (BWT) as mean ± sample standard deviation over three seeds for ImageBind (IB), LanguageBind (LB), and EBind (EB). *Macro* averages the three benchmarks within each backbone. Bold indicates the best result in each column within a block. KeepLoRA is not applicable to EBind (Appendix C).

Backbone	Method	K600 (20 sess.)		VGGSound (40 sess.)		SSW60 (10 sess.)		Macro	
		FAA↑	BWT↑	FAA↑	BWT↑	FAA↑	BWT↑	FAA↑	BWT↑
IB	SeqLoRA	28.59 ± 2.74	−60.43 ± 2.88	18.66 ± 1.43	−69.98 ± 1.20	24.73 ± 2.43	−45.69 ± 2.65	23.99	−58.70
	O-LoRA	44.74 ± 0.48	−41.34 ± 0.62	42.27 ± 0.40	−40.58 ± 0.33	28.35 ± 1.07	−41.70 ± 1.03	38.45	−41.21
	SD-LoRA	28.79 ± 2.81	−60.39 ± 2.92	18.67 ± 1.48	−70.01 ± 1.25	24.73 ± 2.54	−45.69 ± 2.76	24.06	−58.70
	InfLoRA	65.90 ± 0.08	−17.26 ± 0.15	58.04 ± 0.19	−21.91 ± 0.31	58.20 ± 0.13	−10.21 ± 0.15	60.72	−16.46
	KeepLoRA	50.12 ± 0.08	−38.15 ± 0.05	49.79 ± 0.04	−34.82 ± 0.04	46.01 ± 0.64	−22.13 ± 1.05	48.64	−31.70
	BiCLoRA	**71.74 ± 0.01**	**−8.73 ± 0.02**	**63.39 ± 0.06**	**−11.21 ± 0.16**	**61.67 ± 0.21**	**−4.96 ± 0.32**	**65.60**	**−8.30**
LB	SeqLoRA	46.66 ± 1.70	−48.24 ± 1.60	17.97 ± 0.56	−74.27 ± 0.57	32.43 ± 1.55	−49.59 ± 1.99	32.35	−57.37
	O-LoRA	64.38 ± 1.31	−27.52 ± 0.97	49.04 ± 0.91	−38.91 ± 1.32	39.14 ± 1.46	−41.73 ± 0.20	50.85	−36.05
	SD-LoRA	47.20 ± 1.76	−47.73 ± 1.71	17.91 ± 0.63	−74.34 ± 0.61	32.51 ± 1.62	−49.43 ± 2.15	32.54	−57.17
	InfLoRA	76.19 ± 0.08	−11.61 ± 0.07	64.05 ± 0.28	−20.70 ± 0.42	69.07 ± 0.11	−8.97 ± 0.19	69.77	−13.76
	KeepLoRA	64.94 ± 0.20	−27.97 ± 0.23	58.48 ± 0.09	−30.52 ± 0.10	57.09 ± 0.56	−23.44 ± 0.28	60.17	−27.31
	BiCLoRA	**79.71 ± 0.05**	**−6.97 ± 0.04**	**69.02 ± 0.04**	**−11.01 ± 0.04**	**69.94 ± 0.10**	**−7.92 ± 0.27**	**72.89**	**−8.63**
EB	SeqLoRA	58.95 ± 4.55	−36.50 ± 4.60	24.29 ± 2.28	−64.95 ± 2.28	63.19 ± 1.19	−26.22 ± 1.04	48.81	−42.55
	O-LoRA	72.74 ± 0.65	−19.81 ± 0.85	58.54 ± 0.21	−24.77 ± 0.88	68.77 ± 0.86	−19.14 ± 1.35	66.68	−21.24
	SD-LoRA	59.17 ± 4.72	−36.28 ± 4.78	24.51 ± 2.36	−64.77 ± 2.33	63.22 ± 1.19	−26.21 ± 1.02	48.97	−42.42
	InfLoRA	81.84 ± 0.11	−8.96 ± 0.13	64.78 ± 0.09	−15.41 ± 0.05	**83.61 ± 0.26**	−3.45 ± 0.28	76.74	−9.28
	BiCLoRA	**84.25 ± 0.04**	**−5.67 ± 0.06**	**67.64 ± 0.08**	**−9.94 ± 0.04**	82.51 ± 0.05	**−3.05 ± 0.12**	**78.13**	**−6.22**

Table 3: **Ablation of BiCLoRA's two designs.** Disabling Design 1 replaces the history-safe input basis with the leading second-moment directions of the raw inputs. Disabling Design 2 replaces the frozen output basis with a trainable $p \times r$ map. The Design-1-only variant therefore represents input-side protection with bounded history. The final two rows evaluate alternative bases under kernel-only training. ‡ denotes results available only for ImageBind and LanguageBind.

Design		K600 (20 sess.)		VGGSound (40 sess.)		SSW60 (10 sess.)		Macro	
Design 1	Design 2	FAA↑	BWT↑	FAA↑	BWT↑	FAA↑	BWT↑	FAA↑	BWT↑
✓	✓	78.57	−7.12	66.68	−10.72	71.38	−5.31	72.21	−7.72
✓	–	74.64	−12.61	62.29	−19.34	70.30	−7.55	69.08	−13.17
–	✓	77.12	−10.44	65.73	−15.59	68.32	−10.71	70.39	−12.25
–	–	39.51	−54.26	23.68	−66.67	33.06	−49.33	32.08	−56.75
✓	random	78.26	−7.82	65.34	−13.82	70.36	−5.23	71.32	−8.96
SVD of W_m	SVD of W_m ‡	74.62	−10.20	65.70	−13.09	64.46	−7.46	–	–

Ablation study. Table 3 evaluates the two constraints across all nine settings. Without either design, second-moment input directions and a freely trained output factor lead to severe forgetting (32.08 macro FAA and -56.75 BWT). Design 1 alone protects the historical input subspace and raises performance to 69.08 FAA and -13.17 BWT, but leaves output-side alignment unconstrained. Design 2 alone preserves cross-modal alignment (Fig. 4) and reaches 70.39 FAA and -12.25 BWT, but does not protect historical input directions. Combining both achieves 72.21 FAA and -7.72 BWT, confirming that the two constraints address complementary failure modes. Adding Design 2 to the input-protected variant improves FAA by 3.13 points and BWT by 5.45 points while reducing trainable adaptation parameters from $16{,}384$ to 128. Replacing the data-derived output basis with a random orthonormal basis reduces FAA by 0.89 points and BWT by 1.24 points (Table 3): freezing limits alignment drift, and the session-derived basis adds accuracy on top. Results are stable across history, rank and text-weight budgets (Tables 6 and 7).

Visualization of shared cross-modal geometry. Fig. 4 tracks first-session representations on ImageBind–SSW60 under the four design combinations. Without either constraint, representation drift reaches 0.56 and audio-video similarity falls to -0.34. The history-safe input basis alone reduces drift to 0.38, but alignment remains negative at -0.13 because the output factor is unconstrained. The frozen output basis alone preserves positive alignment at 0.17, but drift remains 0.28. Combining both constraints limits drift to 0.02 and maintains an alignment of 0.25, close to 0.32

Figure 4: **Ablation on ImageBind–SSW60. (a)** Drift of first-session joint representations from the pretrained model. **(b)** Their audio-video alignment. Mean ± s.d. over three seeds.

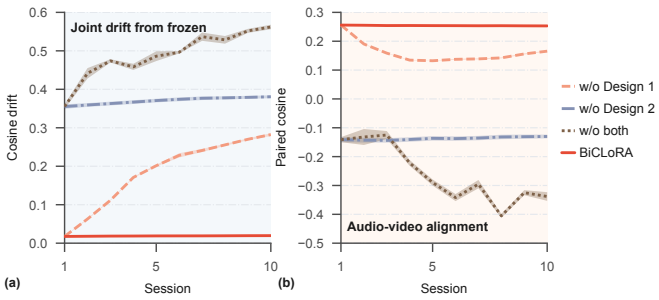


Table 4: **Per-session cost at the ImageBind site.** Trainable adaptation parameters and persistent state over T sessions, excluding the temporary classifier (sizes after the 40 VGGSound sessions; SD-LoRA: session mean).

Method	Trainable	Persistent state
SeqLoRA	32,768	none
O-LoRA	32,768	grows with T
SD-LoRA	32,807	grows with T
InfLoRA	16,384	constant (512 KiB)
KeepLoRA	16,384	grows with T (3.6 MiB)
BiCLoRA	**128**	constant (512 KiB)

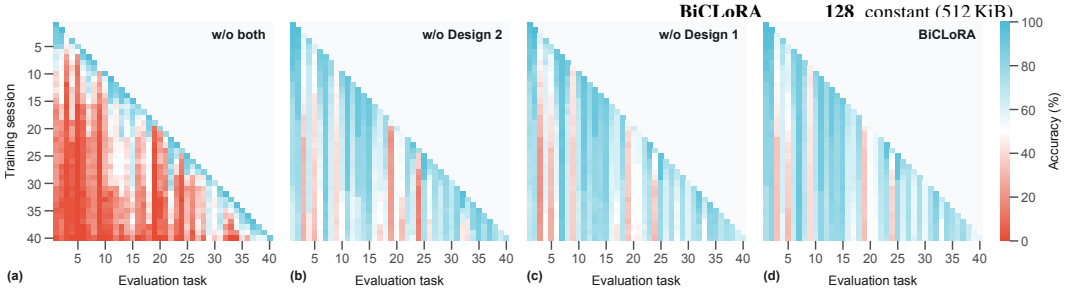


Figure 5: **Session-wise accuracy matrices on ImageBind–VGGSound (40 sessions, seed 1234).** Columns are the evaluated session and rows the training stage. Panels: **(a)** without both designs, **(b)** without Design 2, **(c)** without Design 1, and **(d)** BiCLoRA.

under the frozen model. The t-SNE view in Fig. 8 corroborates this result: SeqLoRA and InfLoRA separate paired audio and video features, whereas BiCLoRA preserves their interleaved geometry.

Class-level retention. Fig. 5 presents session-wise accuracy matrices for the 40-session ImageBind–VGGSound stream under the four design combinations in Table 3. Without either constraint, nearly all earlier sessions are lost. Either constraint alone improves retention but still forgets many classes introduced midstream, whereas BiCLoRA largely preserves their initial accuracy. Table 8 shows that the failures with Design 1 alone occur after successful acquisition: five representative classes learned at 95–100% accuracy decline to 0–35% by the end of the stream. Their samples are reassigned to related classes introduced later, including 95% of railroad-car clips to "train wheels squealing" and 95% of bowling-impact clips to "striking bowling." Adding the frozen output basis raises final accuracy on these classes to 55–100% and substantially reduces such reassignment.

5 Conclusion

This work establishes exemplar-free AVT-CL as a practical setting for continually extending pretrained multimodal representations without retaining past observations. We show that low-rank adaptation fails through two complementary mechanisms: historical interference on the input side and alignment drift on the output side. BiCLoRA translates this diagnosis into a bilateral update that combines a bounded-state, history-safe input basis with a frozen output basis, leaving only a small coupling kernel to be learned and consolidated. Experiments across three pretrained AVT backbones and three benchmarks demonstrate consistent improvements in accuracy and retention, while controlled ablations confirm that both constraints are necessary. These findings highlight a broader principle for continual multimodal adaptation: protecting historical inputs alone is insufficient to preserve the geometry that connects modalities. Our study deliberately isolates representation adaptation under disjoint concept streams, class-name supervision, and a common prediction interface. This controlled setting provides a foundation for extending bilateral constraints to adaptive history compression, modality-aware rank allocation, and internal backbone layers. Further directions include overlapping or evolving concepts, missing modalities, broader distribution shifts, and the joint evolution of representations and prediction interfaces.

Reproducibility Statement

The supplementary material provides the implementation of BiCLoRA and all baselines, the protocols of the three benchmarks, and the configuration of every experiment. It also includes the per-run records (configuration, seed, and accuracy matrix) and a script that recomputes every number in Tables 2–8 from these records.

AI Use Statement

AI-assisted tools were used only for improving writing clarity, grammar checking, and minor editing during manuscript preparation. The authors are solely responsible for the content, including the research methodology, experimental results, and conclusions.

References

Jim Broadbent, Felix Cohen, Frederik Hvilshøj, Eric Landau, and Eren Sasoglu. Ebind: a practical approach to space binding. *arXiv preprint arXiv:2511.14229*, 2025.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725. IEEE, 2020.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190. IEEE, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 3777–3786. IEEE, 2025.

Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in Neural Information Processing Systems*, 37:129146–129186, 2024.

Borui Kang, Lei Wang, Zhiping Wu, Tao Feng, Yawen Li, Yang Gao, and Wenbin Li. Dynamic multi-layer null space projection for vision-language continual learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 2077–2086. IEEE, 2025.

Dawid Kopiczko, Tijmen Blankevoort, and Yuki Asano. Vera: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 6815–6835, 2024.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647. IEEE, 2024.

Yan-Shuo Liang, Jia-Rui Chen, and Wu-Jun Li. Gated integration of low-rank adaptation for continual learning of large language models. *Advances in Neural Information Processing Systems*, 38:68659–68689, 2026.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*, volume 2025, pp. 46461–46477, 2025a.

Yuyang Liu, Qiuhe Hong, Linlan Huang, Alexandra Gomez-Villa, Dipam Goswami, Tiantian Peng, Xialei Liu, Joost van de Weijer, and Yonghong Tian. Continual learning for vlms: A survey and taxonomy beyond forgetting. *arXiv preprint arXiv:2508.04227*, 2025b.

Mao-Lin Luo, Zi-Hao Zhou, Tong Wei, and Min-Ling Zhang. Lada: Scalable label-specific clip adapter for continual learning. *arXiv preprint arXiv:2505.23271*, 2025.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. Keeplora: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 136017–136040, 2026.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

Shentong Mo, Weiguo Pian, and Yapeng Tian. Class-incremental grouping network for continual audio-visual learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 7754–7764. IEEE, 2023.

Tiantian Peng, Yuyang Liu, Shuo Yang, Qiuhe Hong, and YongHong Tian. Gnsp: Gradient null space projection for preserving cross-modal alignment in vlms continual learning. *arXiv preprint arXiv:2507.19839*, 2025.

Weiguo Pian, Shentong Mo, Yunhui Guo, and Yapeng Tian. Audio-visual class-incremental learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 7765–7777. IEEE, 2023.

Kaustubh Ponkshe, Raghav Singhal, Eduard Gorbunov, Alexey Tumanov, Samuel Horvath, and Praneeth Vepakomma. Initialization using update approximation is a silver bullet for extremely efficient low-rank fine-tuning. *arXiv preprint arXiv:2411.19557*, 2024.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset. In *European Conference on Computer Vision*, pp. 271–289. Springer, 2022.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 10658–10671, 2023.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. In *International Conference on Learning Representations*, volume 2025, pp. 58337–58353, 2025.

Xianghu Yue, Xueyi Zhang, Yiming Chen, Chengwei Zhang, Mingrui Lao, Huiping Zhuang, Xinyuan Qian, and Haizhou Li. Mmal: Multi-modal analytic learning for exemplar-free audio-visual class incremental tasks. In *Proceedings of the ACM International Conference on Multimedia*, pp. 2428–2437, 2024.

Huan Zhang, Shuyu Dong, Yujin Zheng, Dingwen Wang, Shenghua Fan, and Fan Lyu. Subspace alignment for clip-based continual learning via canonical correlation analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32257–32266, 2026.

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 *International Conference on Learning Representations*, volume 2024, pp. 9588–9608, 2024.

APPENDIX CONTENTS

A Related Work **12**

B Transfer Study of Vision-Language CL Methods **13**

B.1 Experimental Setup 13

B.2 Experimental Results 13

C Baseline Implementations **14**

D Additional Experimental Results **15**

A RELATED WORK

Audio-visual continual learning. Audio-visual CL studies the joint acquisition and retention of auditory and visual knowledge over evolving data streams. AV-CIL preserves cross-modal associations through similarity constraints and attention distillation, while CIGN organizes multimodal representations using learnable class tokens. Both rely on exemplar replay (Pian et al., 2023; Mo et al., 2023). MMAL enables exemplar-free learning through recursive analytic updates (Yue et al., 2024). These methods establish the importance of retaining knowledge across perceptual modalities, but operate within an audio-video space and do not exploit the linguistic semantics encoded by pretrained tri-modal models. AVT-CL instead uses text as training-time semantic guidance under an exemplar-free protocol, while inference relies only on audio and video. Continual adaptation must therefore acquire new audio-visual knowledge without disrupting the alignment established among all three modalities.

Continual low-rank and subspace adaptation. Parameter-efficient CL commonly builds on LoRA, which represents adaptation as low-rank updates to frozen pretrained weights (Hu et al., 2021). Subspace-based methods mitigate interference by projecting gradients away from historical activation directions, enforcing orthogonality across task-specific factors, or fixing the input side of an update (Saha et al., 2021; Wang et al., 2023; Liang & Li, 2024; Luo et al., 2026). Subspace-constrained LoRA has also been applied to continual learning of CLIP and LLaVA (Luo et al., 2026) and of large language models, where task branches are combined by learned gates (Liang et al., 2026). These approaches protect which inputs an update may respond to, while leaving its output factor trainable. Complementary methods such as VeRA reduce trainable parameters by freezing randomly initialized bases (Kopiczko et al., 2024). LoRA-SB (Ponkshe et al., 2024) trains only an $r \times r$ core between two frozen bases and shows that this parameterization approximates full fine-tuning when the bases come from the first full-model gradient. Neither accounts for earlier sessions. BiCLoRA combines subspace protection with kernel-only training, but constructs both bases from the continual stream: historical projection inputs determine where an update may act, while current-session representations determine where the resulting changes may lie.

Continual adaptation of multimodal representations. Vision-language CL adapts pretrained representations through modules inserted within the encoders or lightweight components learned over frozen features (Jha et al., 2024; Luo et al., 2025). Recent methods also constrain adaptation to preserve image-text geometry (Zhang et al., 2026). Keeping cross-modal alignment intact under continual adaptation has so far been studied mainly for vision–language models. C-CLIP (Liu et al., 2025a) studies image–text continual learning while retaining zero-shot ability; MG-CLIP (Huang et al., 2025) finds that the modality gap tracks how much pre-trained knowledge survives fine-tuning; GNSP (Peng et al., 2025) combines null-space gradient projection with an alignment-preservation loss; and DMNSP (Kang et al., 2025) treats the visual and textual branches asymmetrically after observing that the visual branch drifts more. A recent survey names cross-modal feature drift as a failure specific to multimodal models (Liu et al., 2025b). This literature generally assumes a single perceptual branch paired with text, which often serves as the classifier or remains available at

inference. AVT-CL instead contains two separately parameterized perceptual branches, optimized jointly, that must remain aligned with each other and with the frozen text space, while prediction is based on their fused representation without test-time text. Updating either branch can therefore alter the common coordinate system used by all modalities. BiCLoRA addresses this setting at the final audio and video projections, using text only for training-time supervision and fixed, data-dependent bases to constrain changes within the shared space.

B Transfer Study of Vision-Language CL Methods

Vision-language CL generally assumes a single perceptual branch anchored by text. AVT-CL instead involves two perceptual branches that are adapted jointly and must remain aligned with a frozen text space. To examine whether existing methods transfer to this setting, we evaluate CLAP4CLIP (Jha et al., 2024) and LADA (Luo et al., 2025) on VGGSound and SSW60 with all three backbones, each against a zero-adaptation control. We first describe the experimental setup (Sec. B.1) and then report the results (Sec. B.2).

B.1 Experimental Setup

Experimental matrix. The study comprises three conditions (no adaptation, CLAP4CLIP, and LADA), three backbones, two benchmarks, and three seeds $\{1234, 2025, 3407\}$, yielding 54 runs. We additionally conduct six forward-only probes, one for each backbone-benchmark setting. All runs use the same cached features, data splits, session orders, and prompt templates as the main experiments. Across methods, only the adaptation rule and prediction head vary.

Method implementations. CLAP4CLIP and LADA are implemented as feature-space methods that apply a residual transformation to the fused audio-video representation, $j \mapsto j + \mathrm{Adapter}(j)$. CLAP4CLIP uses a shared two-layer residual MLP with a zero-initialized final layer. LADA uses a shared MLP adapter with a drift penalty equal to 0.05 times the cosine displacement that the adapter induces on previous class prototypes.

Prediction heads and matched controls. We retain each method's prescribed prediction interface when applicable. CLAP4CLIP uses class-text prototypes, and LADA uses the same interface because it does not define a native classifier for the AVT setting. Frozen text prototypes are transformed by a trainable $p \times p$ projection initialized to the identity, with a trainable scale applied to the resulting similarity logits. The adapter and prediction head are jointly optimized over the current-session classes for one epoch using cross-entropy, batch size 64, and learning rate 10^{-3}, corresponding to 8 and 6 steps per session on VGGSound and SSW60. For seed 1234, the training loss decreases by at least 61% over each run. Each method is paired with an identical zero-adaptation control that performs no gradient updates. Across all 18 controls, the optimizer step count is zero, results are identical across seeds, and predictions change only when the candidate label set expands. Their nonzero forgetting therefore reflects label-space expansion alone. Accordingly, Sec. B.2 supports within-head comparisons against matched controls and is not directly comparable with Table 2, which uses the common analytic classifier.

Evaluation and diagnostic probes. In addition to final average accuracy, Table 5 and Fig. 6b report forgetting, $\mathrm{Fgt} = \frac{1}{T-1}\sum_{k=1}^{T-1}\left(\max_{k\le\ell\le T} A_{k,\ell} - A_{k,T}\right)$, which measures the average decline of earlier sessions from their best to final accuracy. For the zero-adaptation control, whose accuracy can decrease only as the label space expands, $\mathrm{Fgt} = -\mathrm{BWT}$. To localize adaptation effects by modality, Fig. 6c re-evaluates each trained checkpoint using either the normalized audio or video representation in place of the fused query, without retraining. The forward-only probe additionally evaluates fused and branch-specific zero-shot accuracy using frozen test representations and class-text prototypes; it performs no gradient updates and never accesses the training split.

B.2 Experimental Results

Vision-language CL does not transfer reliably. Among the 12 method-backbone-benchmark settings, 10 underperform their corresponding zero-adaptation controls, and the remaining two stay

Table 5: **Vision-language CL methods often underperform matched zero-adaptation controls in AVT-CL.** Results use each method's text-prototype classifier and report means over three seeds. Δ denotes the change from the matched control. Bold indicates degradation exceeding 1.5 times the pooled standard deviation.

	VGGSound (40 sess.)			SSW60 (10 sess.)		
	IB	LB	EB	IB	LB	EB
no adaptation, FAA	48.83	57.83	59.55	56.93	54.13	85.84
no adaptation, Fgt	10.95	11.69	9.92	8.48	7.73	3.23
ΔFAA, CLAP4CLIP	**−6.16**	**−2.28**	**−9.60**	**−20.58**	1.55	**−17.68**
ΔFAA, LADA	**−4.96**	**−2.47**	**−8.54**	**−19.33**	2.08	**−20.20**
Fgt after adapting (mean of 2)	22.38	19.15	20.76	10.99	7.52	7.32
ΔFAA, audio-only query (mean of 2)	−0.91	−4.85	−5.25	1.21	0.48	−0.26
ΔFAA, video-only query (mean of 2)	−7.29	−3.62	−5.88	−13.19	−3.95	−12.65

within seed variation. As reported in Table 5 and summarized in Fig. 6, macro final average accuracy decreases by 8.90 to 9.12 points across the two methods, while mean Fgt increases from 8.67 to 14.68. Adaptation strategies developed for vision-language CL are therefore not reliably beneficial when transferred to an aligned audio-video-text space.

Degradation concentrates in the stronger branch. A probe of the frozen representations shows that the video-over-audio zero-shot accuracy gap ranges from 7 to 86 points across the six backbone-benchmark combinations (right axis of Fig. 6c). Branch-wise evaluation of the adapted models shows that this gap predicts where the loss falls: the frozen video-over-audio gap is negatively correlated with the video-over-audio difference in adaptation gains (Spearman $\rho = -0.77$; Fig. 6c). Adapting the fused representation without protecting existing directions therefore damages most the information the pretrained space already encodes well, consistent with the historical interference identified in Sec. 2.2.

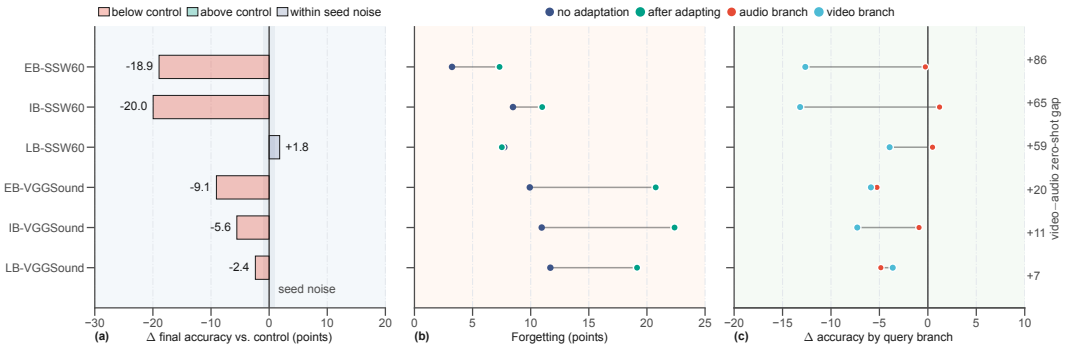


Figure 6: **Vision-language CL transfer frequently degrades AVT-CL performance.** The six backbone-benchmark settings are ordered by their frozen video-over-audio accuracy gap, with results averaged across two methods. **(a)** Change in final average accuracy from the matched zero-adaptation control. Bar styles indicate changes exceeding 1.5 times the pooled standard deviation, and shading denotes ± 1 percentage point. **(b)** Forgetting with and without adaptation. **(c)** Audio-only and video-only accuracy changes. Per-method results are reported in Table 5.

C Baseline Implementations

All low-rank baselines use the adaptation sites defined in Sec. 4.1 and share the same pretrained checkpoints, session data, class orders, optimization budgets, and evaluation protocol. We summarize their implementations under this common AVT-CL interface below.

Low-rank baselines. **SeqLoRA** trains a rank-8 factor pair for each modality in every session and merges the resulting update into the corresponding projection. **O-LoRA**, in a hard-projection

variant of its orthogonality constraint, additionally projects the current input-side factor onto the orthogonal complement of the row spaces spanned by previously stored factors. **SD-LoRA** retains the direction of each previous session update and learns a scalar magnitude for each retained update. **InfLoRA** designs its input basis from the current session's inputs projected off the retained history and trains a free $p \times r$ output map. In the present implementation the retained history is the same rank-64 bounded eigensummary that BiCLoRA uses rather than the growing threshold-selected subspace of the original method, so that the comparison with BiCLoRA isolates the output side of the update. **KeepLoRA** protects a growing principal input subspace: it is initialized with the leading singular directions of the pretrained projection weight (cumulative energy threshold 0.85) and, after each session, extended with the residual principal directions of the current-session projection inputs (threshold 0.99). At the start of each session, both factors are initialized by a randomized SVD of the mean weight gradient projected off this subspace, the initial product is subtracted from the weight so that the effective map is unchanged, the input factor is fixed and only the output factor is trained; the update is merged after the session. We use rank 8 and $\alpha/r = 1$, as for the other LoRA baselines, and the same thresholds for the audio and visual branches. KeepLoRA is not applicable to EBind, whose audio adaptation weight is a zero-initialized residual around a frozen nonlinear projector and whose video interface is an added identity map; neither has a pretrained weight spectrum from which to derive the protected subspace.

Common analytic classifier. All methods in Table 2 are evaluated with the same online analytic classifier on the joint feature, with the text branch frozen. The classifier maps the joint feature through a $2{,}048$-dimensional blockwise-orthogonal random projection, constructed through QR decomposition with sign correction and scaling by the number of blocks, and solves a ridge regression with coefficient 10^3 from accumulated sufficient statistics. The projection is fixed across methods within each backbone-benchmark setting.

Optimization. All projection-adaptation methods are trained for one epoch per session with batch size 1, using AdamW with a learning rate of 5×10^{-4} for the adaptation parameters and 10^{-3} for the temporary classifier, no weight decay, gradient-norm clipping at 1.0, and bfloat16 autocast.

Adaptation interface. Each projection is applied to every clip of a sample. On ImageBind, clip outputs are ℓ_2-normalized, averaged over the valid clips, and renormalized; on LanguageBind, they are averaged and normalized; on EBind, audio clips are normalized before averaging and video clips are not. The history summary uses the per-clip projection inputs, with the clips of a sample sharing a total weight of one, whereas the output basis uses the pooled sample representations. For EBind audio, $W_{a,0} = 0$ is the initial residual rather than a pretrained projection, and z^a in Eq. 8 is the full audio representation $\mathrm{Norm}(\mathrm{MLP}(h^a) + W_a h^a)$; because the MLP is fixed, the update still changes the representation by $\Delta W_a h^a$ before normalization. For LanguageBind, the class-text representation is the normalized average of the normalized outputs of its two text encoders.

D Additional Experimental Results

Full ablation across all nine settings. Table 7 extends the ablations in Table 3 to all nine backbone-benchmark settings and includes sweeps over the history rank Q_m, total rank R, and text weight λ_{text}. The two designs remain complementary on Kinetics-600: removing Design 1 reduces FAA by 0.8–2.3 points, while removing Design 2 costs 2.4–5.8 points. BiCLoRA is otherwise robust to its main hyperparameters. Across all Q_m and R values, macro FAA ranges from 71.94 to 72.47, at least 2.8 points above the strongest baseline. Increasing Q_m modestly improves BWT, from -8.44 at $Q_m = 16$ to -7.51 at $Q_m = 128$, at the cost of additional state. The default $R = 16$ is fixed to match the LoRA baselines without dataset-specific tuning; $R = 8$ and $R = 32$ achieve comparable macro FAA of 72.17 and 72.47, respectively. Performance is also stable across $\lambda_{\text{text}} \in \{0.01, 0.05, 0.2\}$, yielding 72.21–72.24 macro FAA; class text additionally initializes the temporary classifier g_t independently of this weight. Table 6 further shows that retaining the exact, untruncated history is neither efficient nor beneficial. It requires approximately $12\times$ the default state and produces the lowest FAA and BWT among the history budgets (71.72 and -8.52), as the protected subspace eventually exhausts the available input directions. Across ImageBind and LanguageBind, 381 of 420 sessions retain fewer safe directions than the requested rank, including 198 with none available, whereas the bounded summary always leaves $d_m - Q_m$ candidate directions.

Table 6: **Effect of the history budget Q_m.** We report the combined persistent state of the audio and video summaries (KiB), macro accuracy, and backward transfer across the nine settings. A rank-Q_m summary stores $Q_m(d_m + 1)$ floats per modality. The total is identical across backbones because they share the same $d_a + d_v$. The exact summary retains all supported eigendirections and requires up to $d_m(d_m + 1)$ floats per modality, with its state measured on EBind.

History summary	persistent state (audio + video)	FAA↑ (9 settings)	BWT↑ (9 settings)
Q_m=16	128 KiB	71.94	−8.44
Q_m=32	256 KiB	72.12	−8.04
Q_m=64 (default)	512 KiB	72.21	−7.72
Q_m=128	1025 KiB	72.16	−7.51
exact history (no truncation)	6010 KiB	71.72	−8.52

Table 7: **Full ablation across all nine backbone-benchmark settings.** FAA is reported as the mean over three seeds (%). *Macro* gives FAA and BWT across all nine settings. ‡ denotes results available only for ImageBind and LanguageBind, since EBind's identity-initialized visual projection has no preferred singular subspace.

Variant	K600 (20 sess.)			VGGSound (40 sess.)			SSW60 (10 sess.)			Macro	
	IB	LB	EB	IB	LB	EB	IB	LB	EB	FAA↑	BWT↑
w/o Design 1 (second-moment input directions)	69.44	78.51	83.40	61.48	68.77	66.96	57.66	65.34	81.95	70.39	−12.25
w/o Design 2 (trainable $p \times r$ output map)	65.90	76.19	81.84	58.04	64.05	64.78	58.20	69.07	83.61	69.08	−13.17
w/o both designs	25.42	39.00	54.10	16.57	21.66	32.81	21.44	22.95	54.78	32.08	−56.75
both bases frozen, taken from the SVD of the weight‡	69.84	79.40	–	62.84	68.57	–	60.24	68.68	–	–	–
output basis U^{out} replaced by a random orthonormal basis	71.30	79.36	84.11	62.17	66.84	67.03	59.26	69.60	82.23	71.32	−8.96
history rank Q_m=16	71.05	79.49	83.97	63.30	68.97	67.31	60.65	70.07	82.67	71.94	−8.44
history rank Q_m=32	71.52	79.58	84.19	63.42	68.99	67.55	60.98	70.21	82.66	72.12	−8.04
history rank Q_m=128	71.89	79.86	84.34	63.44	68.79	67.57	61.29	69.70	82.52	72.16	−7.51
exact history (no truncation)	71.78	79.41	84.45	63.28	69.12	67.72	60.61	66.48	82.60	71.72	−8.52
total rank R=8	71.95	80.06	84.45	63.17	68.75	67.46	61.27	69.91	82.48	72.17	−7.43
total rank R=32	71.91	79.93	84.30	63.70	69.29	67.29	61.91	70.96	82.91	72.47	−7.96
text weight λ_{text}=0.01	71.71	79.71	84.23	63.37	68.95	67.67	61.69	70.02	82.61	72.22	−7.71
text weight λ_{text}=0.2	71.77	79.74	84.24	63.40	68.96	67.71	61.77	70.03	82.53	72.24	−7.68
Full BiCLoRA	71.74	79.71	84.25	63.39	69.02	67.64	61.67	69.94	82.51	72.21	−7.72

Finally, freezing both bases but deriving them from the SVD of the projection weights preserves the geometry comparably well, yet trails BiCLoRA by 1.0 macro FAA point over the six applicable settings (68.26 versus 69.25), confirming that accuracy depends on selecting data-supported bases in addition to freezing them.

Output-basis control: freezing vs. selecting directions. Removing Design 2 changes both the output parameterization and its number of trainable parameters: the output factor becomes trainable, increasing the adaptation size from r^2 to $p\,r$. To isolate direction selection, the random-basis variant in Table 7 retains BiCLoRA's history-safe input basis, rank, 8×8 kernel, and training budget, but replaces U_m^{out} with a random orthonormal basis redrawn each session. This variant preserves the geometry at least as well as BiCLoRA, with first-session drift of 0.007 and alignment of 0.31 on ImageBind–SSW60, confirming the importance of freezing the output subspace. However, it underperforms BiCLoRA in all nine settings, reaching 71.32 versus 72.21 macro FAA and -8.96 versus -7.72 BWT. Relative to the variant without Design 2, freezing a random output basis recovers approximately 2.2 FAA points; selecting this basis from the current-session output second-moment matrix contributes a further 0.9 points. Thus, both restricting the output degrees of freedom and selecting data-supported directions contribute to performance.

Forgetting cases. Table 8 reports the five ImageBind–VGGSound classes with the largest retention gap between BiCLoRA and the variant without Design 2. Although every variant initially learns these classes almost perfectly, the trainable output factor later redirects them toward semantically related classes introduced in subsequent sessions. For example, 95% of railroad-car clips are ultimately classified as "train wheels squealing," and 95% of bowling-impact clips as "striking bowling." Input-side protection alone cannot prevent this reassignment because new classes may

still be written into output directions occupied by earlier ones. Freezing the output basis enables BiCLoRA to retain 55–100% accuracy on these classes. The same pattern occurs on ImageBind–SSW60, where, relative to the variant without Design 2, BiCLoRA improves final accuracy from 43% to 73% for Eastern Phoebe, from 67% to 90% for Baltimore Oriole, and from 70% to 85% for American Goldfinch.

Table 8: **Representative forgetting cases on ImageBind–VGGSound (seed 1234).** For each class, top-1 accuracy (%) is reported immediately after its introduction and at the end of the stream. The bottom block identifies the dominant class to which its samples are reassigned without Design 2 and reports the corresponding assignment rate under each variant.

Class	railroad car, train wagon	race car, auto racing	bowling impact	hail	playing bongo
Session introduced	31	31	3	12	23
SeqLoRA	100 → 5	95 → 0	100 → 0	100 → 0	95 → 0
w/o Design 2	100 → 0	100 → 10	100 → 5	100 → 35	95 → 0
w/o Design 1	100 → 70	90 → 75	100 → 45	100 → **100**	95 → 45
BiCLoRA	100 → **90**	90 → **85**	100 → **70**	100 → **100**	95 → **55**
Absorbing class (session)	train wheels squealing (37)	skidding (34)	striking bowling (35)	raining (32)	playing congas (24)
share, w/o Design 2	95%	80%	95%	35%	65%
share, w/o Design 1	20%	15%	55%	0%	30%
share, BiCLoRA	**0%**	**5%**	**30%**	**0%**	**20%**

Additional visualizations. Fig. 7 extends the drift analysis to the 40-session ImageBind–VGGSound stream, while Fig. 8 visualizes the first-session features after completing ImageBind–SSW60. Because each t-SNE panel is embedded independently, comparisons concern only its internal geometry. Lines connect paired audio and video features from the same clip. SeqLoRA and InfLoRA separate the two modalities into distinct clusters, whereas BiCLoRA preserves their pairing as in the frozen model. The annotated joint-feature drift and audio-video similarity follow Sec. 4.2 and use seed 1234. Across the VGGSound and SSW60 settings of all three backbones, SeqLoRA and InfLoRA induce historical-feature drift of 0.22–0.64, while BiCLoRA limits it to at most 0.07. On SSW60, the baselines reduce audio-video similarity below zero (-0.13 to -0.37), whereas BiCLoRA remains within 0.08 of the frozen model. On VGGSound, BiCLoRA keeps the alignment positive (0.18–0.39), while SeqLoRA and InfLoRA can reduce it to -0.09 and 0.00, respectively. KeepLoRA, available on ImageBind and LanguageBind, shows a drift of 0.41–0.44 in these settings and an SSW60 similarity of -0.14 and -0.22.

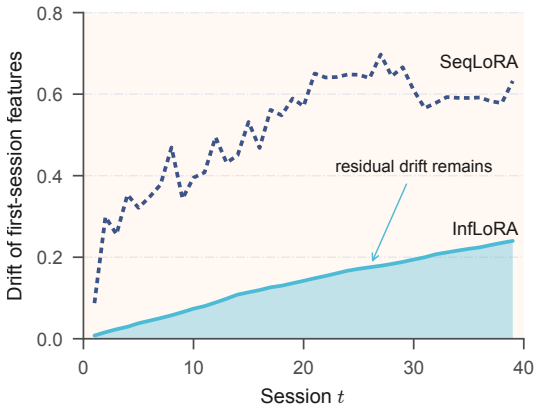


Figure 7: **Residual drift over a long stream (ImageBind–VGGSound, 40 sessions).** Cosine drift of first-session visual features under SeqLoRA and input-protected adaptation. Audio exhibits the same trend at a smaller magnitude. The interference amplification in Fig. 2a, $d\,\mathrm{tr}(\Delta_t C \Delta_t^\top)/\|\Delta_t\|_F^2$ with $C = P\Lambda P^\top/\mathrm{tr}(\Lambda)$, is zero by construction for InfLoRA and BiCLoRA, yet residual drift remains because the bounded history summary is incomplete.

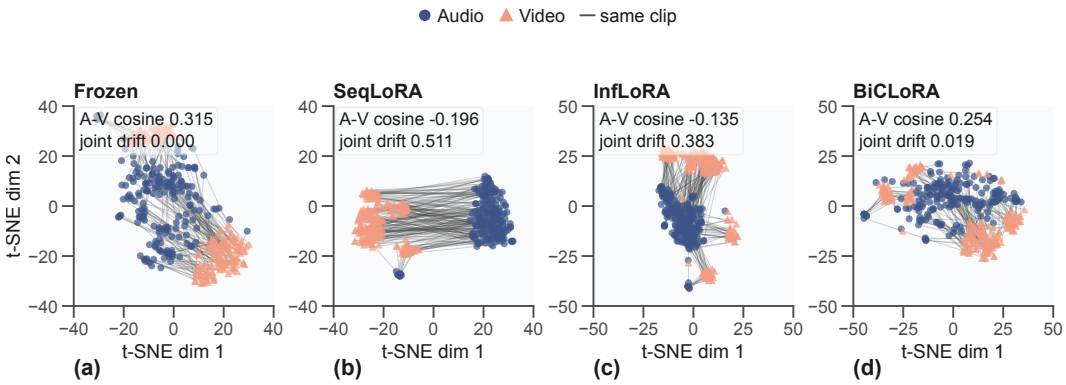


Figure 8: **Cross-modal pairing of old classes after the full ImageBind–SSW60 stream (seed 1234).** t-SNE visualizations show the audio and video features of the six first-session classes at $t = T$ under the frozen model, SeqLoRA, InfLoRA, and BiCLoRA. Colors and markers distinguish modalities, while lines connect paired features from the same clip. Annotations report mean audio-video similarity and joint-feature drift from the frozen model.