THE MORE, THE MERRIER: CONTRASTIVE FUSION FOR HIGHER-ORDER MULTIMODAL ALIGNMENT

Supplementary Material

Supplementary Material

This supplementary document provides additional details and results that complement the main paper. It is organized as follows:

- **Experimental Details.** We present comprehensive descriptions of model configurations, hyperparameters, training procedures, and evaluation protocols to ensure reproducibility.
- **Additional Ablation Studies.** We include supplementary ablation experiments that highlight additional aspects of this work, examining how different model components and design choices influence overall performance.
- **Additional Results.** This section includes further quantitative results that extend and support the findings presented in the main paper.
- **Datasets.** Additional information regarding the datasets used and our new dataset Bird-MML.

A. Experimental Details

A.1. Experiments on Bird-MML

Dataset preprocessing For the audio modality, Each audio clip was truncated or zero-padded to a duration of 10 seconds and resampled to 22.05 kHz. Mel spectrograms were computed using a window size of 1024 and a hop length of 512, resulting in 128 Mel frequency bands. Each spectrogram was treated as a single-channel image to ensure compatibility with a ResNet-50 backbone.

For the image modality, Images were resized to 224×224 pixels and normalized. No data augmentation was applied.

Model Architecture. The multimodal architecture employed is composed of three modality-specific encoders:

- **Text encoder:** BERT transformer.
- **Image encoder:** ResNet-50.
- **Audio encoder:** ResNet-50 applied to Mel-spectrogram inputs.

Each encoder outputs a 512-dimensional embedding. A two-layer multilayer perceptron (MLP) was used for multimodal feature fusion. All models and baselines were trained using identical hyperparameters, without any additional hyperparameter search or tuning.

Training used the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. A batch size of 64 was used throughout the training process. The learning rate followed a cosine annealing schedule with a minimum of 1×10^{-6}. The fusion loss coefficient was set to $\lambda = 0.5$.

A.2. MOSI, MUStARD, and UR-FUNNY

For these datasets, all models were trained for 100 epochs, producing 256-dimensional representations. We used a batch size of 128 and a learning rate of 5×10^{-5}. Following prior work [20], each modality encoder was implemented as a 5-layer Transformer with 5 attention heads.

Fusion was achieved through a 2-layer MLP that combined unimodal embeddings into a single fused representation. Dataset splits followed the official `MultiBench` configuration for training, validation, and testing.

A.3. AVMNIST

For the text modality, we used a BERT transformer. Image, spectrogram, and satellite-view inputs were encoded using ResNet-18 backbones initialized from ImageNet and fully trained. Spectrograms provided by the `MultiBench` dataloaders were used without modification.

Embeddings had a 256-dimensional size. Dataset splits again followed the `MultiBench` protocol.

Training Setup. All models were trained from scratch for 30 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 256. The best model was selected based on validation loss. Each experiment was repeated with five random seeds, and we report the mean and standard deviation across runs. For our model, MLP-based fusion was used with a balancing parameter $\lambda = 0.5$ between pairwise and fusion losses.

A.4. XOR Experiment

For the XOR experiment, we mapped all modalities to embeddings of size 128. The dataset consisted of 10,000 training samples and 5,000 test samples. Models were trained for 50 epochs with a batch size of 512 and a learning rate of 1×10^{-4}.

Each encoder was implemented as a 2-layer MLP, and we used a balanced fusion coefficient of $\lambda = 0.5$.

B. Additional Ablation Studies

B.1. Effect of the λ Parameter

As shown in Table 7, performance is not overly sensitive to λ. Accuracy remains stable while recall varies smoothly,

indicating a controlled trade-off with minimal tuning required. Identifying a principled way to set or eliminate λ is an important direction.

Table 7. Effect of λ across datasets (mean $\pm$ std). Accuracy (%) remains stable while λ trades off recall. Best values in bold.

Dataset	λ	Acc. (%)	$R^{1\to1}_{mean@10}$	$R^{2\to1}_{mean@10}$
MOSI	0.2	**67.7** $\pm$ 1.7	19.5 $\pm$ 1.0	**20.1** $\pm$ 1.9
	0.5	66.7 $\pm$ 2.1	**19.5** $\pm$ 1.5	19.6 $\pm$ 2.0
	0.8	65.8 $\pm$ 2.0	13.5 $\pm$ 2.2	15.6 $\pm$ 1.6
UR-Funny	0.2	64.2 $\pm$ 0.7	**8.0** $\pm$ 0.4	**14.0** $\pm$ 0.9
	0.5	**64.9** $\pm$ 1.0	7.4 $\pm$ 0.2	13.6 $\pm$ 0.6
	0.8	64.6 $\pm$ 0.7	6.9 $\pm$ 0.3	13.2 $\pm$ 0.7
MUSTARD	0.2	59.7 $\pm$ 3.2	42.1 $\pm$ 2.2	57.1 $\pm$ 2.5
	0.5	**64.1** $\pm$ 2.5	**45.3** $\pm$ 2.6	**62.6** $\pm$ 3.1
	0.8	61.1 $\pm$ 3.4	41.0 $\pm$ 1.0	59.7 $\pm$ 1.1

Regarding optimal λ values the results in Table 8 highlight the influence of the λ parameter, which balances the contribution of modality-specific and shared representations. These λ values were selected from the Pareto front obtained by jointly optimizing the two retrieval objectives - mean Recall@10 for 1→1 and 2→1 tasks. We observe that moderate λ values (0.1–0.5) yield the best trade-off between the two objectives, and that the optimal λ varies across datasets, indicating optimal λ depends on data distribution.

Table 8. Mean Recall@10 results for 1→1 and 2→1 tasks across datasets.

Dataset	λ	**Mean R@10 (1→1)**	**Mean R@10 (2→1)**
MOSI	0.1	21.21	22.12
UR-FUNNY	0.4	8.06	14.69
MUSTARD	0.5	45.29	62.56

B.2. Embedding Dimensionality Analysis on the XOR Task

In this experiment, we evaluate how the embedding dimensionality affects each method's ability to solve the XOR problem for the case of $\hat{p} = 1$. Figure 6 illustrates the mean accuracy as a function of the embedding dimension. As provided, SYMILE successfully learns to solve the XOR task even with a minimal embedding dimensionality of 8, achieving perfect accuracy thereafter. Our proposed method (CONFU) also manages to solve XOR but requires a larger embedding dimensionality of 64 to reach convergence. In contrast, both TRIANGLE and GRAM fail to solve the task even when the embedding dimensionality is increased up to 1024, implying that the bottleneck lies in their loss formulations rather than in representational capacity.

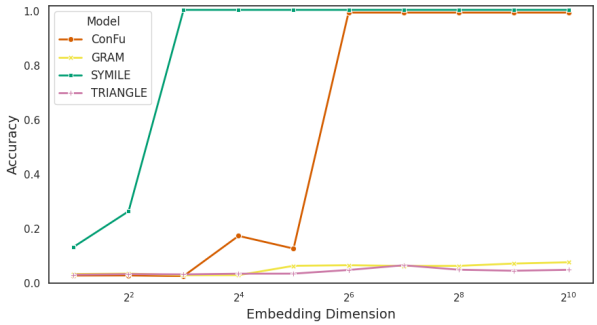


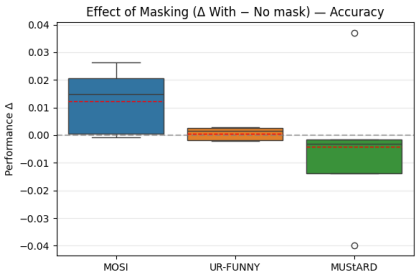
Figure 6. Accuracy as a function of embedding dimension for $\hat{p} = 1$. SYMILE achieves perfect accuracy from dimension 8 onwards, while CONFU requires a dimensionality of 64 to converge. TRIANGLE and GRAM fail to solve the XOR task even at 1024 dimensions.

B.3. Tackling modality shortcuts

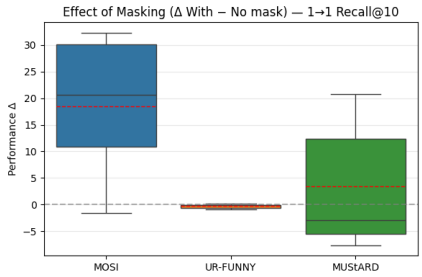
Motivated by the hypothesis that the fusion network may exploit shortcut strategies, we investigated the effect of applying dropout at the feature level, specifically, before concatenation and subsequent processing by the fusion MLP. A potential shortcut can arise when the pairwise alignment losses between modalities become sufficiently low, allowing the fusion module to rely predominantly on one modality while suppressing the contribution of the other. In such cases, the network may learn to transmit information through the most predictive modality instead of developing genuinely integrated multimodal representations. Introducing feature-level masking may counteract this behavior by preventing the model from depending exclusively on any single modality and encouraging more balanced multimodal learning.

To examine this effect, we compared model performance with and without masking for multiple λ values and mask ratios. For each dataset and task, we computed the performance difference Δ = (With mask) − (No mask) across runs and visualized these distributions using boxplots.

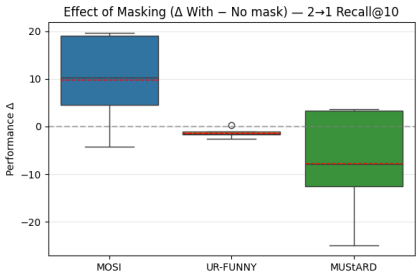
As shown in Figure 7, the MOSI dataset consistently benefits from masking across all tasks, classification as well as 1→1 and 2→1 retrieval, indicating that feature-level dropout can improve representation robustness and reduce reliance on modality-specific cues. In contrast, the UR-FUNNY dataset exhibits minimal changes, suggesting that masking has limited impact on its multimodal dynamics, while MUStARD experiences performance degradation, likely due to its stronger dependence on precise cross-modal alignment. Moreover, as shown in Table 9, the use of masking in ConFu yields significant performance gains across all query→target retrieval directions on the MOSI dataset. Overall, these findings suggest that the impact of masking is dataset-dependent: it tends to improve perfor-



(a) Performance difference (Δ With - No mask) for the **Accuracy** task. Each box shows variability across λ values for different datasets.



(b) Performance difference for **1→1 retrieval (Recall@10)**. Positive values indicate that masking improves unimodal retrieval.



(c) Performance difference for **2→1 retrieval (Recall@10)**. Each box aggregates runs with different λ values and mask ratios.

Figure 7. Effect of the masking ratio on multimodal performance for the MLP fusion model. Each subplot reports the performance change (Δ = With mask − No mask) across datasets and tasks, computed over multiple λ and mask ratio values. These results illustrate how both hyperparameters jointly influence model behavior for classification and retrieval tasks.

mance when redundant cross-modal information could otherwise enable shortcut learning, but may hinder it when successful alignment relies on using the full representational capacity of each modality.

Table 9. Recall@10 (%) on MOSI with and without masking (mask ratio = 0.3, $\lambda = 0.5$). Δ indicates the absolute improvement.

Target	Query(s)	Without Mask	With Mask	Δ
M1	**M2**	21.02 ± 1.65	**25.10** ± **3.51**	+4.08
	M3	16.41 ± 2.32	**20.41** ± **1.51**	+4.00
	M23	16.73 ± 1.84	**22.24** ± **3.16**	+5.51
M2	**M1**	19.18 ± 1.41	**25.83** ± **2.98**	+6.65
	M3	21.02 ± 2.40	**22.77** ± **0.96**	+1.75
	M13	21.63 ± 1.89	**24.58** ± **2.33**	+2.95
M3	**M1**	16.06 ± 1.74	**23.35** ± **1.61**	+7.29
	M2	23.47 ± 2.69	**24.40** ± **1.91**	+0.93
	M12	20.50 ± 2.87	**24.34** ± **3.02**	+3.84

B.4. Noise experiments

We investigate the robustness of ConFu under controlled noise-induced distribution shifts applied to either the image or audio modality. Gaussian noise was added at varying severities with standard deviations of 0.05, 0.1, and 0.15 for images, and 0.1, 0.2, and 0.3 for audio. The corresponding SNR values are estimated from the data.

The results in Table 10 show that ConFu exhibits improved robustness to noise compared to competing baselines. Notably, GRAM and TRIANGLE appear largely unaffected when noise is added to audio. However, this behavior is likely attributable to their effective disregard of the audio modality. In contrast, both methods degrade substantially when noise is introduced to the image modality, performing even worse than unimodal TriCLIP.

Table 10. Accuracy (%) under modality-specific noise-induced distribution shift. Noise was applied to individual modalities at test time with varying SNR levels. A = Audio, V = Vision, A+V = Audio–Visual fusion. Best performance for each degradation type and level is reported in bold.

	10dB SNR		15dB SNR		20dB SNR	
Method	**A deg.**	**V deg.**	**A deg.**	**V deg.**	**A deg.**	**V deg.**
Tri-CLIP (V)	69.0	5.5	69.0	13.2	69.0	35.4
Tri-CLIP (A)	26.0	**31.1**	29.1	31.1	30.3	31.1
Symile [27]	58.4	21.2	59.5	27.9	60.0	40.9
GRAM [5]	58.9	4.0	58.3	8.94	58.1	25.9
TRIANGLE [6]	63.9	3.4	64.0	7.24	64.0	26.5
ConFu	**71.2**	30.2	**71.4**	**33.1**	**71.5**	**45.4**

Overall, ConFu consistently outperforms all baselines across noise levels, including unimodal variants that do not suffer degradation. Its performance trails unimodal TriCLIP-audio slightly only in the extreme case where the visual modality becomes heavily corrupted. A key observation is that Symile, GRAM, and TRIANGLE deteriorate markedly under noise, whereas ConFu incurs only minimal performance loss, particularly when noise is added to audio, where the accuracy drop is negligible. For comparison, TriCLIP experiences roughly a 5% accuracy decline at an audio SNR of 10 dB, while ConFu drops by only $\sim$0.2%.

B.5. Discussion on Multimodal Competition

In Fig. 8, we present a detailed per-class modality overlap analysis for the SSW60 dataset in the zero-shot classification setting. This experiment visualizes, in the form of a heatmap, the distribution of overlap categories for each label separately. Specifically, each cell indicates the percentage of samples belonging to a given class that fall into one of the following overlap categories: *Audiovisual Only*, *Vision Only*, *Audio Only*, *Audiovisual & Vision*, *Audiovisual & Audio*, *Vision & Audio*, *All*, or *None*. These categories

capture whether a sample is correctly classified by only one modality, by a specific combination of modalities, by all modalities, or not correctly classified at all.

Higher values in a cell correspond to a larger proportion of samples for which that modality (or modalities) yields a correct prediction. This visualization allows us to inspect class-specific modality behavior, highlighting both complementary and competing contributions across modalities.

We observe that label 31 is exclusively correctly classified by the *Audiovisual Only* modality, suggesting that neither vision nor audio alone provides sufficient information to recognize this class. This reinforces the need for complementary cross-modal information in order to correctly classify certain bird species.

C. Additional Results

C.1. Few shot adaptation on SSW60, VB100, CUB200 complete results

In this section, we present the complete results of our few-shot adaptation experiments on the SSW60, VB100, and CUB200 [32] datasets. The following figures include all baseline methods for comparison and illustrate adaptation performance across varying number of shot settings. These results provide a comprehensive view of how each method generalizes to limited-data scenarios and adapts to novel classes.

SSW60. In the multi-frame setting, where mean embeddings over 8 video frames are used, our fusion-based method achieves the best performance. It is important to note that competing approaches do not support multimodal representations, and therefore only unimodal results are available for comparison. In the vision modality, most methods perform similarly, with Symile showing a slight drop. In contrast, in the audio modality, TRIANGLE and GRAM underperform significantly, indicating that their generated audio representations are not informative. In the single-frame setting, **ConFu**'s audiovisual representations outperform all unimodal variants by a substantial margin.

VB100. In VB100, audiovisual fusion performs worse than vision-only models. This aligns with our earlier discussion: the audio modality is largely uninformative and even distracting for this dataset, as reflected by the low performance of audio-only baselines.

CUB200. For CUB200, we evaluate on classes completely unseen during pretraining. Symile achieves the strongest performance, while our method (**ConFu**) ranks second.

D. Datasets

D.1. Bird-MML Dataset

Dataset Construction.

To construct our dataset, we combined data from three sources: **iNaturalist** for images, **Xeno-Canto** for audio, and **Wikipedia** for textual and class-level grounding. For every bird class present in the VB100 or SSW60 datasets, we collected corresponding images (from iNaturalist), audio recordings (from Xeno-Canto), and Wikipedia articles.

Audio Processing Audio samples were segmented into 10-second clips (or zero-padded if shorter). We used the tags provided in Xeno-Canto as semantic grounding for the final captions.

Image Captioning Each image from iNaturalist was processed with `InstructBLIP2` to generate visual descriptions, which served as grounding for the final caption.
Model: `Salesforce/instructblip-flan-t5-xl`
Prompt:

```
Describe the bird's colors,
size, and shapes.
```

Wikipedia Text Processing For each class, we sampled random sections from the corresponding Wikipedia page. These sections were summarized by an LLM to produce textual grounding, generating 100 captions per class. This allowed us to capture diverse factual descriptions of each species.
Model: `google/gemma-2-2b-it`
Prompt:

```
You are a naturalist assistant.
Summarize the following text
into one coherent caption.
Be concise and factual.
{section_text}
Caption:
```

Caption Combination & Triplet Construction In the final step, we matched images, audio clips, and Wikipedia summaries into triplets. To produce a unified caption grounded in all three modalities (image, class, audio), we fed the image caption, audio tags, and a randomly selected Wikipedia caption into an LLM.
Prompt:

```
You are a naturalist assistant.
Combine the following
information into a concise
caption (10--20 words).
```

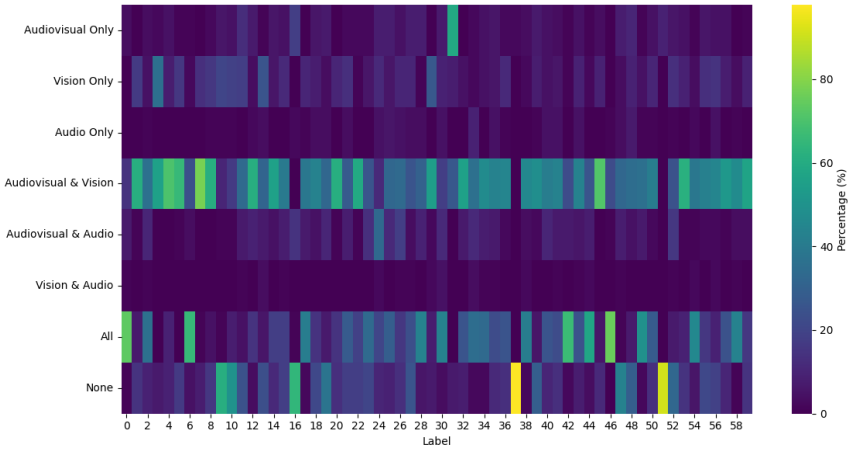


Figure 8. Per-class modality overlap analysis on the SSW60 dataset in the zero-shot classification setting. Each cell shows the percentage of samples within a class that fall into a given overlap category: *Audiovisual Only*, *Vision Only*, *Audio Only*, *Audiovisual & Vision*, *Audiovisual & Audio*, *Vision & Audio*, *All*, or *None*. Higher values indicate a larger fraction of samples for which that modality (or combination) produces correct predictions. This visualization highlights class-specific complementarity and competition between modalities.

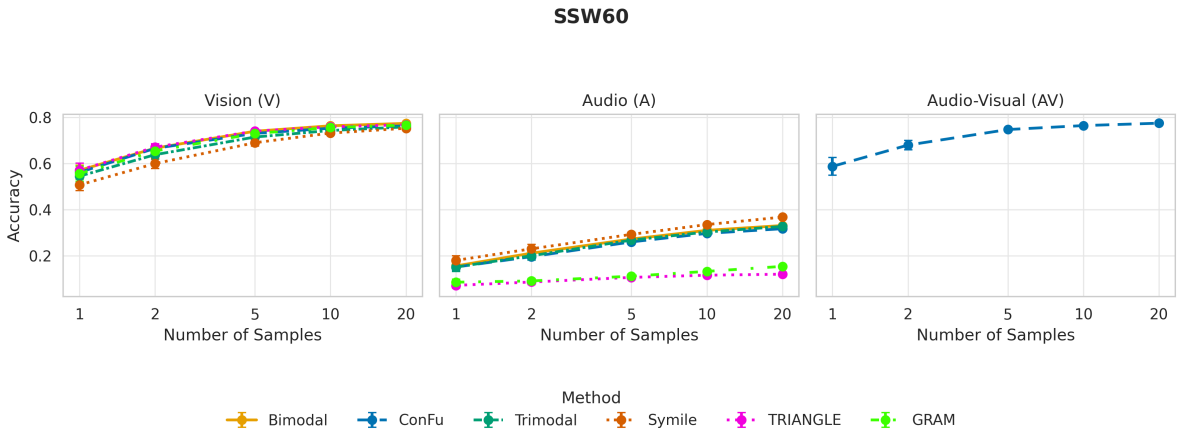


Figure 9. Few-shot linear probing results for SSW60 [30]. Performance is shown as the number of labeled examples increases. Prediction is done in the 8-frame sampling setting (multi-frame).

```
Include information about the
bird's appearance and sound if
available.
Be factual and informative.

Image caption: {image info}
Species caption: {textual info}
Sound tags: {tags}

Caption:
```

This process results in coherent multimodal captions grounded jointly in visual appearance, species knowledge, and acoustic characteristics.

D.2. Affective Computing Benchmarks

MultiBench [19] is a collection of benchmarking datasets designed to evaluate multimodal representation learning across a wide range of modalities and domains. In our experiments, we used the MOSI, UR-FUNNY, and MUStARD datasets. MultiBench provides pre-extracted features and predefined dataset splits, both of which we adopt in our setup.

MOSI is a multimodal sentiment analysis dataset consisting of 2,199 annotated YouTube video clips. UR-FUNNY contains 16,514 samples extracted from TED Talks and focuses on humor detection in spoken language. MUStARD includes 690 video clips from popular TV shows and is designed for multimodal sarcasm detection.

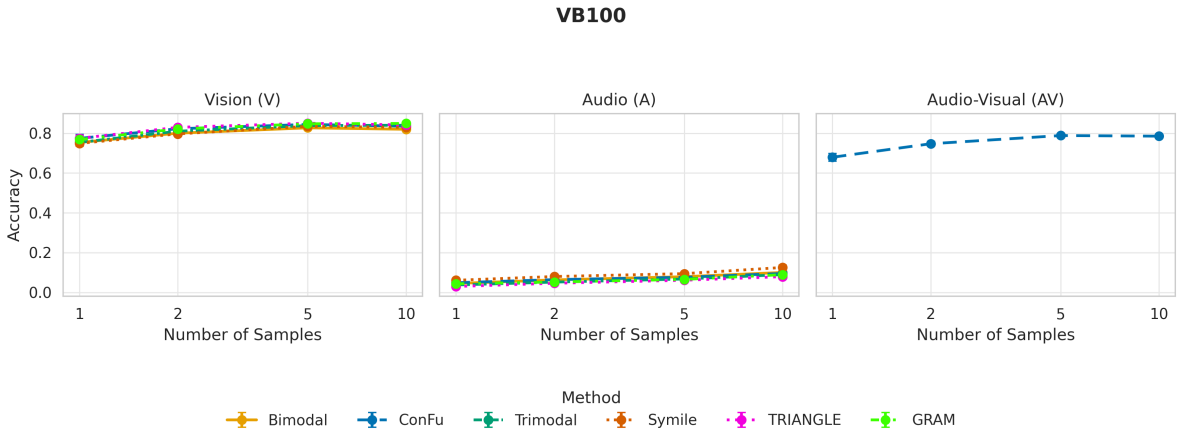


Figure 10. Few-shot linear probing results for VB100 [9]. Performance is shown as the number of labeled examples increases. Prediction is done in the 8-frame sampling setting (multi-frame).

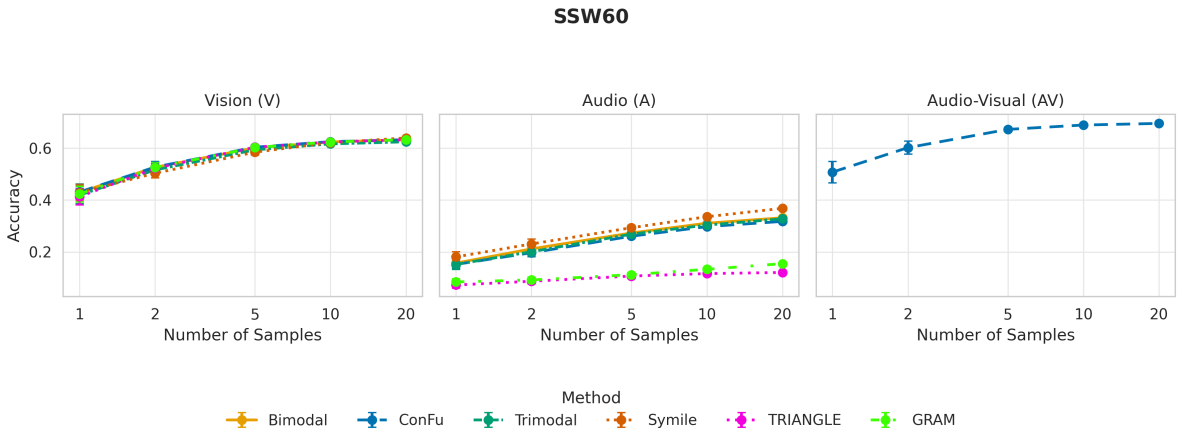


Figure 11. Few-shot linear probing results for SSW60 [30]. Performance is shown as the number of labeled examples increases.Prediction is done in the single frame setting.

D.3. AV-MNIST

The AV-MNIST dataset is a multimodal benchmark that pairs degraded visual features with audio spectrograms to evaluate multimodal representation learning. The visual modality consists of 28×28 MNIST digits that have been PCA-projected, retaining only 25% of the total variance. This dimensionality reduction is applied intentionally to weaken the visual signal and encourage effective fusion across modalities.

The audio modality comprises 112×112 spectrograms generated from the Free Spoken Digits Dataset [15], with additive background noise sampled from ESC-50 [25] to increase variability and realism.

The dataset contains 55000 training samples and 10000 test samples. In our experiments, we use the pre-extracted audio spectrograms distributed by MultiBench [19].

E. Generalization to M Modalities

For the generalization of our framework to M modalities, we extend the objective in Eq. 7 by summing InfoNCE losses over all relevant modality subsets. Specifically, we include all disjoint subset pairs

$$\mathcal{L}_M = \sum_{\substack{S_i, S_j \subseteq \{1,\ldots,M\} \\ S_i \cap S_j = \emptyset,\ S_i, S_j \neq \emptyset}} \widehat{\mathcal{L}}^{(S_i,S_j)}_{\text{InfoNCE}}, \tag{15}$$

which correspond to mutual information terms $I(X_{S_i}; X_{S_j})$ estimated via InfoNCE bounds. This formulation provides a direct extension of our tri-modal framework, allowing contrastive learning to capture both pairwise and higher-order dependencies across arbitrary modality combinations.

To ascertain the above claim, we conducted preliminary experiments in a four-modality setting. This was achieved by splitting the original images into full spatial resolu-

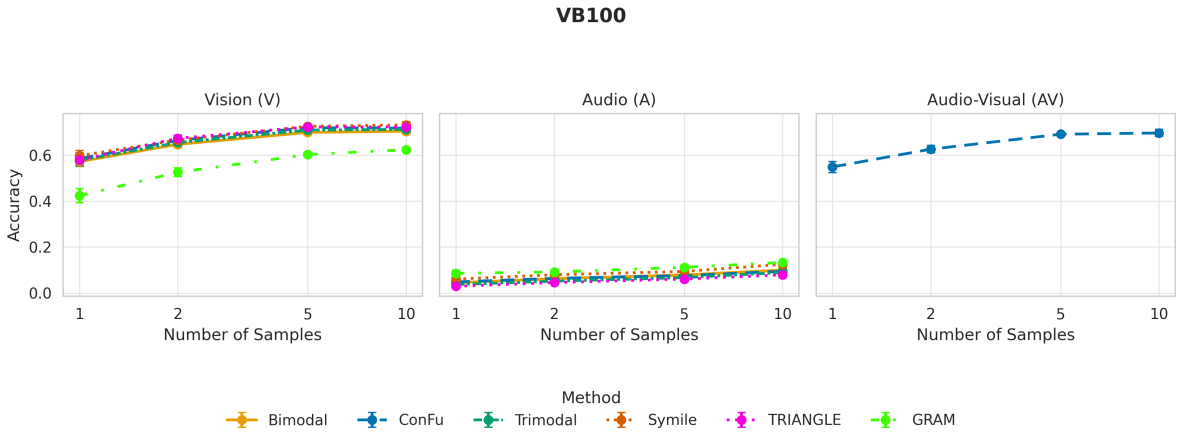


Figure 12. Few-shot linear probing results for VB100 [9]. Performance is shown as the number of labeled examples increases.Prediction is done in the single frame setting.

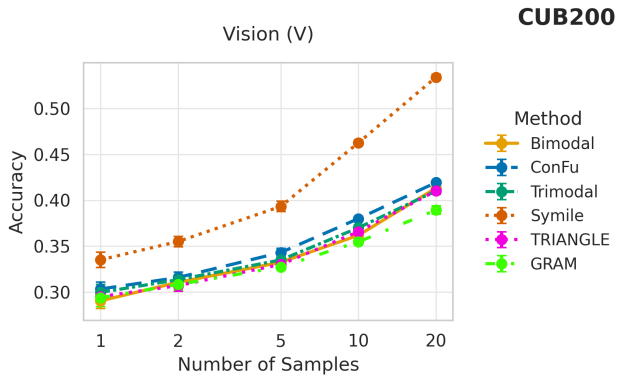


Figure 13. Few-shot linear probing results for CUB200 [32]. Performance is shown as the number of labeled examples increases. Prediction is done in the single frame setting.

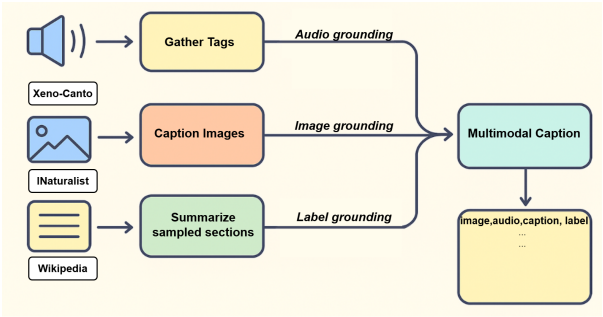


Figure 14. Overview of the multimodal data generation pipeline. Images and audio are collected from public sources (iNaturalist and Xeno-Canto), pseudocaptions are extracted from Wikipedia, and captions are generated via a combination of InstructBLIP2 and Gemma-2-2B-IT. The resulting triplets form the pretraining dataset.

tion monochromatic (grayscale) and low-resolution color (RGB), yielding four modalities alongside audio and text without additional data sources. Although artificial, this scenario emulates the behavior of numerous remote sensing platforms (e.g., Pléiades: 0.5 meter monocromatic and 2 meters RGB) and the associated problem of pansharpening. Experiments on VB100, and SSW60 show that the composite objective converges and consistently improves zero-shot performance over a pairwise-only baseline (Table 11), providing evidence that the approach remains effective as additional modalities are incorporated (with VB100 results reflecting the known weakness of the audio signal).

Table 11. Zero-shot classification accuracy (%) on SSW60 and VB100 datasets. G = Grayscale Vision, LR = Low-Res RGB Vision, A = Audio.

	SSW60 Acc. (%)		VB100 Acc. (%)	
Modality	**Pairwise CLIP**	**ConFu**	**Pairwise CLIP**	**ConFu**
G (grayscale)	50.33	48.46	14.48	15.47
LR (low-res RGB)	58.28	55.22	15.89	16.74
A (audio)	26.67	26.02	3.04	1.77
G + LR	–	59.39	–	18.15
G + A	–	52.20	–	14.05
LR + A	–	57.63	–	15.32
G + LR + A	–	**61.96**	–	**15.32**

Nevertheless, while the formulation in Eq. 15 captures all possible cross-subset dependencies, its combinatorial nature results in a large number of contrastive terms as M increases. In practice, task-specific relaxations of this general objective can substantially reduce computational complexity. For instance 16, when the goal is to retrieve a single target modality, one may restrict the loss to align any subset of the remaining modalities with that target only, i.e., terms of the form $I(X_t; X_S)$. Such relaxations preserve the rep-

resentational alignment relevant to the retrieval task while avoiding the exponential growth in the number of InfoNCE terms, yielding a more tractable yet principled multimodal contrastive objective.

$$\mathcal{L}_{\text{retrieval}} = \sum_{\substack{S_i, S_j \subseteq \{1,\ldots,M\} \\ S_i \cap S_j = \emptyset,\ S_i, S_j \neq \emptyset \\ |S_i| = 1}} \widehat{\mathcal{L}}_{\text{InfoNCE}}^{(S_i, S_j)}. \tag{16}$$