# Efficient and High-Fidelity Omni Modality Retrieval

## Supplementary Material

In this supplementary material, we provide additional details regarding our proposed OmniRet model and the newly introduced ACM benchmark, which complement the findings in the main paper. Specifically, Sec. 7 details the instructions used for data generation, the architecture of the Shared Media Resampler, and the pseudo-code for ASWP. Sec. 8 describes the full statistics of our benchmark and the specifics of the baseline settings employed. Finally, Sec. 9 outlines all datasets and baselines utilized in our experiments, including the chosen hyper-parameters for each training stage. We conclude with detailed quantitative results on all benchmarks and provide additional ablation studies on the proposed modules, including a comparative analysis of the computational costs across different models.

## 7. Additional Method Details

We leverage the instructions introduced in M-BEIR [71] and extend them to cover the specific requirements of our new datasets. We provide a complete list of all instructions used during model training in the code. Note that the first instruction in this file is the one employed during inference and testing.

We provide further details regarding our architecture. The media projectors within OmniRet employ the standard design adopted by LLaVA, utilizing two-layer Multi-Layer Perceptrons with GELU [19] non-linear activations interleaved between them. For the shared media resampler, we utilize a module consisting of two cross-attention blocks. The hidden dimension throughout the resampler is kept consistent and equal to the dimension of both the input and output features. In our experiments, we set the number of learnable latent vectors to $N = 64$. This constraint was imposed by our available computational resources, as increasing $N$ beyond this limit would require a corresponding decrease in batch size and significantly extend the training time. While a larger $N$ may potentially yield improved performance, we found $N = 64$ to be the maximum trainable size within our resource limits. For video processing, we incorporate learnable frame embeddings to capture temporal information. These embeddings are initialized using a standard sinusoidal pattern.

The ASWP layer follows the standard Perceiver architecture for its main components. It consists of two cross-attention blocks where the hidden dimension is consistently maintained to match the input and output dimensionality of the layer. The procedural steps of the ASWP mechanism are formally detailed in Algorithm 1.

**Algorithm 1:** Attention Sliced Wasserstein Pooling

**Data:** Pooled LLM Hidden States: $\mathbf{Z} = \{\mathbf{z}_i \in \mathbb{R}^D\}_{i=1}^S$
**Result:** Embedding vector: $\mathbf{h} \in \mathbb{R}^L$
**Trainable parameters:**
  Sliced Reference Set: $\mathbf{X}' = \{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^S$;
  Slicers (Projectors): $\Theta = \{\theta_l \in \mathbb{R}^D\}_{l=1}^L$;
**for** $l = 1$ *to* $L$ **do**
  Calculate sliced input: $\mathcal{S}_l^{\mathbf{Z}} = \{\mathbf{z}_i\theta_l\}_{i=1}^S \in \mathbb{R}^S$;
  Sort the distributions: $\pi_{\mathbf{Z}} = \text{argsort}(\mathcal{S}_l^{\mathbf{Z}})$ and $\pi_{\mathbf{X}} = \text{argsort}(\mathbf{X}')$;
  Let $\pi_{\mathbf{X}}^{-1}$: indices that permutes the sorted reference set back to the origin;
  Compute per-slice embedding: $\psi_l = \{\mathcal{S}_l^{\mathbf{Z}}[\pi_{\mathbf{Z}}[\pi_{\mathbf{X}}^{-1}[i]] - \mathcal{S}_l^{\mathbf{X}}[i]\}_{i=1}^S \in \mathbb{R}^S$;
**end**
Set of embeddings: $\mathbf{Z}' = \{\psi_l\}_{l=1}^L \in \mathbb{R}^{S\times L}$;
Pooled embedding $\mathbf{h}$ computed via STM;

## 8. Additional Benchmark Details

Table 6. The Statistics of the Audio-Centric Modality Benchmark

| Task | No. Queries | No. Candidates |
|---|---|---|
| A,T→A | 1292 | 4251 |
| A→I | 1292 | 5480 |
| I→A | 1292 | 5480 |
| A→V | 1292 | 5480 |
| V→A | 1292 | 5480 |

The statistics for the five retrieval tasks in the ACM benchmark are presented in Table 6. We used the prompts specified in Table 7 and Table 8 to generate the audio captions and modification texts, respectively. To ensure data quality, a subjective evaluation was conducted using the Qualtrics survey platform to verify the quality of the generated audio captions and modification texts. Fig. 5 illustrates the structure of the survey used for this verification process.

We conduct additional studies to assess the quality of ACM on a 300-sample subset. Since the dataset is generated by Gemini 2.5, we use GPT-4o for evaluation to avoid self-preference bias. For each sample, we score naturalness, fluency, and hallucination on a 1–5 scale. ACM achieves strong average scores of 4.4, 4.1, and 4.5, respectively. On a clean subset of 250 samples (Table 9), where all scores are at least 4, OmniRet maintains results consistent with Table 4, indicating that synthetic artifacts or distribution shifts do not favor our method.

Table 7. Prompt structure to generate audio captions in the Audio-Centric Multimodal Benchmark.

| | |
|---|---|
| **System** | `You are a language assistant that helps to generate audio captions.` |
| **Prompt** | `Help me generate a concise audio caption, which does not exceed 30 words. The generated audio should focus on describing acoustic events within the audio and their temporal relationships. Only generate the caption without explanation.`<br><br>`[AUDIO]` |

Table 8. Prompt structure to generate modification texts in the Audio-Centric Multimodal Benchmark.

| | |
|---|---|
| **System** | `You are a language assistant that helps to generate the modification text between two audio captions.` |
| **Prompt** | `Generate the modification text for the following pair of audio captions:`<br>`Caption 1: [CAPTION 1]`<br>`Caption 2: [CAPTION 2]`<br><br>`<Good example>`<br>`...`<br>`</Good example>`<br><br>`<Bad example>`<br>`...`<br>`</Bad example>`<br><br>`You need to answer the changes from audio 1 to audio 2. The text needs to be concise and details as you can hear the audio, not as you are reading the text. The modification text should not exceed 20 words. You should not add words "details, specific, description, context, clarify, description, and caption" to the text. Also, you should not focus on the sound location, only focus on the acoustic sounds and their temporal relationship. You should focus on the coustic aspect of these captions, which can only be heard, do not focus on the visual aspect which can only be seen, such as "large, small, big, or a specific object/subject name".` |

To comprehensively evaluate the effectiveness of the OmniRet model on the ACM benchmark, we compare it against three strong baselines: a text-to-text retrieval method, CLAP, and ImageBind.

**Text-to-Text Retrieval Baseline**. We benchmark against a standard text-to-text retrieval approach using the Gemma embedding model [68]. For this baseline, we utilize audio and video captions generated by the Qwen Omni2.5 model [76].

- *Composed Audio Retrieval*: The query is formed by con-

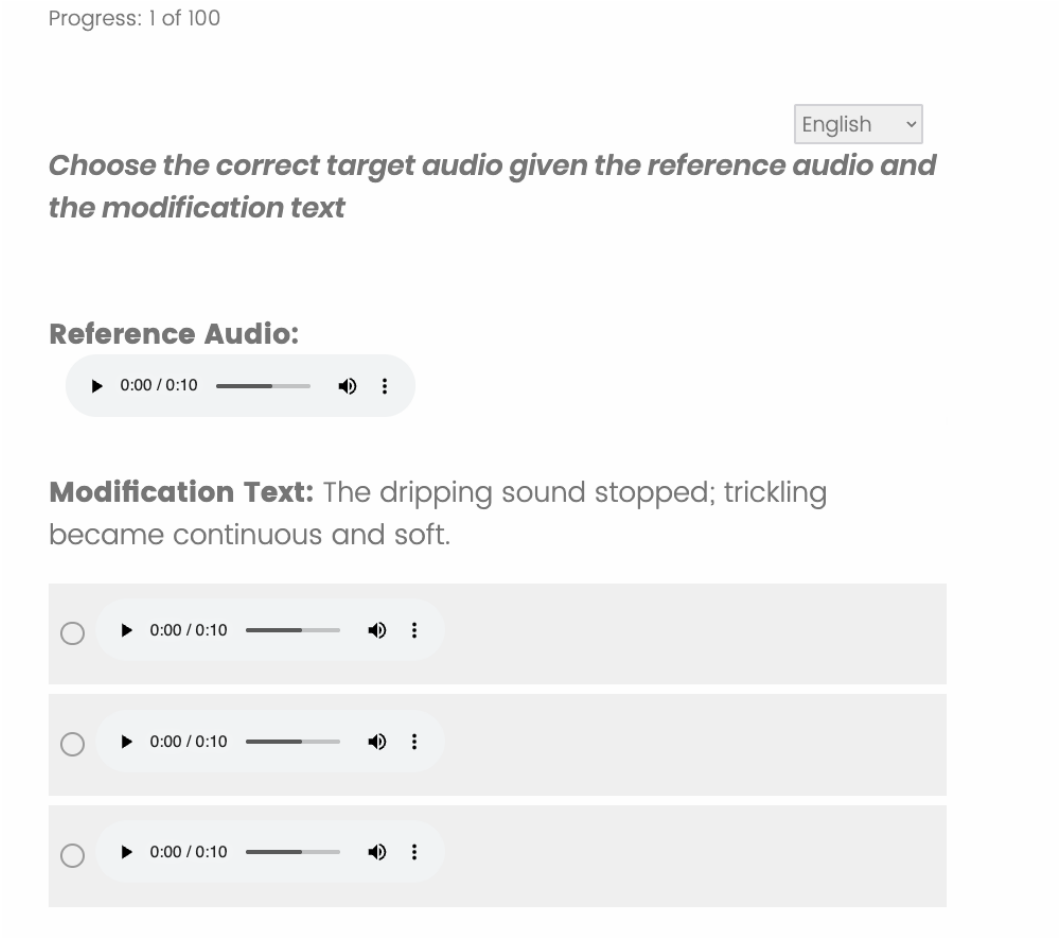


Figure 5. The Qualtric survey for verifying the quality of generated texts in our ACM benchmark.

Table 9. Performance on ACM Clean Subset.

| Model | A,T → A | A → V | V→ A | A→ I | I→ A |
|---|---|---|---|---|---|
| CLAP [12] | 18.4 | - | - | - | - |
| ImageBind [16] | 10.5 | 38.7 | 39.4 | **33.9** | **33.5** |
| OmniRet | **28.2** | 34.2 | 36.7 | 26.0 | 24.0 |

catenating the query audio caption $\mathbf{t}_{\text{cap}}$ with the corresponding modification text $\mathbf{t}_{\text{mod}}$:

$$\mathbf{q} = \mathbf{t}_{\text{cap}} || \mathbf{t}_{\text{mod}} \tag{11}$$

The candidates are the candidate audio captions provided in the benchmark.

- *Audio-Visual Retrieval Tasks*: We directly use the audio and video captions generated by the Qwen Omni2.5 model to perform these retrieval tasks.

**CLAP**. We also benchmark against the CLAP model [74]. We adapt the fusion method from the UniIR framework [71] to handle the composed audio retrieval task. Specifically, the query is modeled as a linear combination of the query audio embedding and the modification text embedding:

$$\mathbf{h}^{\mathbf{q}}_{\text{CLAP}} = \lambda \cdot g_A(\mathbf{a}) + (1-\lambda) \cdot g_T(\mathrm{t}_{\text{mod}}) \tag{12}$$

Here, $g_A(\mathbf{a})$ and $g_T(\mathrm{t}_{\text{mod}})$ represent the embedding vectors from the audio and text encoders of the CLAP model, respectively, with the mixing weight $0 < \lambda < 1$. We set $\lambda = 0.5$ in our experiments to obtain the results reported in Table 4. The candidates are the audio embeddings of the candidate audios, encoded by the CLAP audio encoder. Note that we do not evaluate CLAP on the audio-visual tasks, as the model is not designed to handle visual modalities.

**ImageBind**. Another baseline is the ImageBind model [16], which learns a joint embedding space across six different modalities.

- *Composed Audio Retrieval*: We follow the identical experimental setup described in Eq. (12) with $\lambda = 0.5$.
- *Audio-Visual Retrieval Tasks*: We extract embeddings using ImageBind's audio and visual encoders and compute the cross-modal similarities directly within the shared

embedding space.

**ViT-Lens and OmniBind**. Similar to ImageBind, two recent works, ViT-Lens [34] and OmniBind [70], learn a joint embedding space across multiple unimodal encoders. We follow the same task setting as ImageBind: for composed audio retrieval, we fuse audio and text embeddings, while for audio-visual retrieval, we directly use their joint embeddings.

# 9. Additional Experimental Details

## 9.1. Training Datasets

Table 10 provides the comprehensive details of all datasets used for training and testing. In total, the training set comprises 6.4 million query-candidate pairs, while the testing set involves approximately 300K queries and a large pool of 5.7 million candidates.

## 9.2. Baselines

*CLIP* [57] is a seminal dual-encoder framework trained on a dataset of 400 million image-text pairs using a contrastive language-image pretraining objective. It learns a shared embedding space by maximizing the cosine similarity between matched pairs while minimizing it for mismatched ones, enabling robust zero-shot generalization across various visual tasks.

*SigLIP* [66] introduces a pairwise sigmoid loss for language-image pretraining, replacing the standard softmax normalization used in contrastive learning. This objective decouples the batch size from the loss computation, allowing for greater scalability and more stable training with larger batch sizes, ultimately yielding superior performance on multimodal retrieval benchmarks.

*PE-Core* [3] proposes extracting visual embeddings from intermediate layers of a pretrained vision transformer rather than the final output layer, demonstrating that these deeper, high-quality feature maps provide superior semantic information. The framework is inherently capable of encoding videos by applying its efficient encoder frame-by-frame, ensuring high-fidelity, perceptually rich representations for video understanding and retrieval tasks.

*MMT* [54] is the first work that introduced text-to-audio retrieval tasks on the AudioCaps and Clotho datasets. The MMT framework leverages the dual towers paradigm to perform audio-text retrieval tasks, where the audio encoder is the ResNet18 pretrained on the VGG-Sound dataset and the text encoder is the word2vec model [51]

*CLAP* [74] is a dual-encoder model pretrained with a contrastive learning objective on a large amount of audio-caption pairs, about 630k pairs. Its learned representation demonstrates a strong alignment across audio and text modalities, enabling Zero-shot and supervised audio classification capabilities.

*CLIP*$_{SF}$ [71] is an instruction-guided multimodal retriever which is further finetuned from the pretrained CLIP model [57] on the training set of the M-BEIR dataset. The model adopts two fusion approaches, score-level and feature-level fusions, to handle eight multimodal retrieval tasks.

*MM-Embed* [41] is the first work to introduce an MLLM-based universal multimodal retrieval framework. It finetunes a strong text retrieval LLM together with a pretrained visual encoder using modality-aware hard negatives and MLLMs as zero-shot rerankers to achieve strong performance for composed retrieval tasks.

*VLM2Vec-V1* [26] is a Vision Language model trained with a contrastive learning objective to generate multimodal representation vectors. It demonstrates a strong embedding space and can handle any composition of image and text to perform composed retrieval tasks on the MMEB benchmark.

*VLM2Vec-V2* [27] is the second version of the VLM2Vec-V1 model. It is trained with a flexible and scalable data sampling procedure to achieve a strong generalization across multimodal tasks.

*UniME* [17] is the a novel two-stage framework which adopts MLLMs to learn a universal representative embedding space for multimodal downstream tasks. It proposed two key techniques, textual discriminative knowledge distillation and hard negative enhanced instruction tuning, to achieve robust discriminative and compositional abilities in multimodal retrieval tasks.

*LLaVE* [32] is a universal multimodal LLM-based retriever which adapts hardness-weighted contrastive learning to address the high overlap in similarity distribution between positive and negative pairs. This model shows not only a strong performance in the MMEB benchmark but also good generalization in text-video retrieval in a zero-shot manner.

*B3++* [65] is a novel batch construction pipeline which is able to harvest high-quality batches from a pretrained teacher model for contrastive learning. The retrieval model trained with this pipeline demonstrates superior performance in the MMEB benchmark with a strong teacher model while achieves a good generalization across domains and tasks with weaker teacher models.

*MetaEmbed* [75] is a universal multimodal framework which employs the Matryoshka MultiVector training technique [30] to organize information by granularity across multiple latent vectors. This framework is capable of performing test-time scaling for multimodal retrieval tasks, enabling trade-off between efficiency and effectiveness.

*ColPaliv1.3* [14] is a specialized Vison Language Model trained with late interaction matching objetive for producing high-quality embedding vectors from image of document pages. This model is a simple yet effective approach which outperforms traditional pipelines for document re-

Table 10. Analytics of dataset use to train and test OmniRet.

| Task | Dataset | Training | | Testing | |
|---|---|---|---|---|---|
| | | No. Queries | No. Candidates | No. Queries | No. Candidates |
| I→I | NIGHTS [15] | 15.9K | 15.9K | 2.1K | 40K |
| T→T | WebQA [5] | 16K | 27K | 2.5K | 544K |
| | MSMarco [2] | 100K | 99K | - | - |
| | HotpotQA [77] | 84.5K | 84.5K | - | - |
| | NaturalQuestion [31] | 100K | 100K | - | - |
| | PAQ [35] | 100K | 100K | - | - |
| | StackExchange [62] | 100K | 100K | - | - |
| | NLI [4] | 100K | 100K | - | - |
| | SQuAD [58] | 87.3K | 87.3K | - | - |
| I→T | VisualNews [43] | 100K | 99.9K | 20K | 537.6K |
| | Fashion200K [18] | 15K | 12.2K | 4.9K | 61.7K |
| | MSCOCO [42] | 113K | 543K | 5K | 24.8K |
| | LLaVA-558K [44] | 272K | 272K | - | |
| T→I | VisualNews [43] | 99.9K | 100K | 20K | 542K |
| | Fashion200K [18] | 15K | 48.9K | 1.7K | 202K |
| | MSCOCO [42] | 100K | 72.4K | 24.8K | 5K |
| | LLaVA-558K [44] | 272K | 272K | - | - |
| V→T | TGIF [38] | 78K | 78K | 33K | 33K |
| | Charades [61] | 8K | 8K | 8K | 8K |
| | WebVid2M [1] | 500K | 500K | - | - |
| | PE-Video [3] | 104K | 45K | - | - |
| T→V | TGIF [38] | 78K | 78K | 33K | 33K |
| | Charades [61] | 8K | 8K | 8K | 8K |
| | WebVid2M [1] | 500K | 500K | - | - |
| | PE-Video [3] | 45K | 104K | - | - |
| A→T | AudioCaps [29] | 91.3K | 91.3K | 975 | 4.9K |
| | ClothoV2.1 [11] | 3.8K | 3.8K | 1K | 5.2K |
| | WavText5K [10] | 4.3K | 4.3K | - | - |
| | WavCaps [49] | 403K | 403K | - | - |
| T→A | AudioCaps [29] | 91.3K | 91.3K | 4.9K | 975 |
| | ClothoV2.1 [11] | 3.8K | 3.8K | 5.2K | 1K |
| | WavText5K [10] | 4.3K | 4.3K | - | - |
| | WavCaps [49] | 403K | 403K | - | - |
| T→I,T | WebQA [5] | 17.2K | 17.1K | 2.5K | 403K |
| | EDIS [45] | 25.9K | 66.2K | 3.2K | 1M |
| I,T→T | OVEN [21] | 151K | 54K | 50K | 677K |
| | InfoSeek [7] | 141K | 22.4K | 11.3K | 612K |
| I,T→I | FashionIQ [73] | 16.2K | 16.2K | 6K | 74.4K |
| | CIRR [47] | 26.1K | 15.7K | 4.2K | 21.6K |
| | CC-CoIR [67] | 250K | 250K | - | - |
| | MTCIR [22] | 208K | 208K | - | - |
| I,T→I,T | OVEN [21] | 154K | 32.6K | 14.7K | 335K |
| | InfoSeek [7] | 143K | 17.9K | 17.6K | 488K |
| V,T→V | WebCoVR [67] | 500K | 500K | 2.6K | 2.6K |
| A→V | VGGSound [6] | 192.5K | 192.5K | - | - |
| V→A | VGGSound [6] | 192.5K | 192.5K | - | - |
| **Total** | | 6.4M | 6.4M | 287K | 5.7M |

trieval tasks.

*GME* [81] is a general embedding model which is trained on a large amount of synthetic query-target pairs spanning across diverse multimodal tasks. This modal proposes the fused-modal training techniques to enable universal modality retrieval.

### 9.3. Training Hyper-parameters and Procedure

Table 11 summarizes the complete set of hyperparameters used in the two distinct training stages of our OmniRet.

**Stage 1: Warm-up**. The first training stage focuses on uni-modality and text-binding tasks, excluding all video-related tasks.

- *Data and Batching:* Each batch is constructed to contain samples from 6 tasks spanning 17 unique datasets. This stage utilizes approximately 2 million total samples.
- *Optimization:* We employ a cosine learning rate (LR) scheduler. The LR is warmed up over 100 steps before decreasing to a minimum value of $1 \times 10^{-4}$.
- *Hardware:* All experiments were conducted using 64 NVIDIA RTX 2080 Ti GPUs with mixed precision (FP16) enabled for computational efficiency.

**Stage 2: Fine-tuning**. Stage 2 continues the training process using the model weights initialized from Stage 1, incorporating all tasks, including the video-related ones.

- *Initialization:* Training begins by resuming the model with a constant learning rate of $1 \times 10^{-4}$ for all modules, with no further warm-up applied.
- *LLM Fine-tuning (LoRA):* We fine-tune the LLM components using LoRA, setting the rank to $r = 16$ and the scaling factor $\alpha = 64$.
- *Batch Adjustment:* To increase the diversity and quality of training signals, we double the number of intra-dataset samples per batch. This is achieved by decreasing the number of distinct tasks included in each batch while simultaneously increasing the overall batch size.
- *Stabilization:* Gradient accumulation is applied throughout this stage to ensure training stability between tasks despite the increased batch size.
- *Data Scale:* The model is trained on a substantially larger scale in this stage, using around $18.4$ million total samples.
- *Minimum LR:* The minimum learning rate for all modules is set to zero.
- *Hardware:* We trained on the same system with 128 GPUs.

## 9.4. Extended Quantitative Results

The performance breakdown across all datasets is detailed in Table 15. This table also includes the results achieved by OmniRet after the first training stage, demonstrating the impact of the full training curriculum. To evaluate on the video tasks, video media tokens are initialized with the image media tokens within the Shared Media Resampler.

While the model exhibits large performance improvements after the second stage, particularly across new tasks, we observe only a marginal decrease in performance on the original uni-modality tasks as the model adapts to the expanded task complexity. On the text-binding tasks, our method achieves outstanding performance on audio and video modalities, alongside comparable performance to state-of-the-art methods on all image-text tasks. Furthermore, in composed visual-text tasks, OmniRet achieves the best score on 6 out of 9 datasets, often establishing a substantial margin over the second-best competitor.

Table 11. **Training Hyper-parameters in Stage 1 and Stage 2**.

| | Stage 1 | Stage 2 |
|---|---|---|
| Batch size | 2,048 | 3,072 |
| No. Tasks | 6 | 15 |
| No. Datasets | 17 | 28 |
| No. Dataset/Batch | 6 | 4 |
| No. Intra-Dataset Samples | $\approx 341$ | 768 |
| No. GPUs | 64 | 128 |
| No. iterations | 1,000 | 6,000 |
| No. training samples | $\approx$ 2M | $\approx$ 18.4M |
| Optimizer | AdamW | |
| Visual Projector LR | 5e-4 | 1e-4 |
| Audio Projector LR | 5e-4 | 1e-4 |
| ASWP LR | 5e-4 | 1e-4 |
| LoRA LR | - | 1e-4 |
| LoRA Rank | - | 16 |
| LoRA Alpha | - | 64 |
| No. Grad. Accum. Steps | 1 | 2 |
| LR Scheduler | Cosine | |
| No. Warmup Steps | 100 | 0 |
| Min LR | 1e-4 | 0 |

Table 12 presents a focused comparison of our OmniRet with the VLM2VecV2 baseline on a subset of the MMEBv2 benchmark. We note that direct comparison on image-text tasks may be compromised due to potential misalignment in the training datasets used by the respective models. Crucially, however, on video-text benchmarks, which primarily employ zero-shot evaluation settings, our model consistently outperforms VLM2VecV2 by a very large margin. This strong performance differential effectively validates the effectiveness of our architecture design and specialized training paradigm for robust video information encoding.

## 9.5. Additional Ablation Studies

We perform three additional ablation studies to rigorously demonstrate the effectiveness of our proposed modules and hyperparameter choices.

We first isolate the contribution of our proposed resampler and pooling design against strong baselines. Extending from Table 5, we match the batch size of 1024 (half of the standard size), which is identical to the batch size used for the "Multi-Vector (Late Attention)" setting. As shown in Table 13, our proposed solution, even with this reduced batch size, remains significantly more effective than completely removing the resampler. While the "Multi-Vector" approach performs slightly better under the same batch size, this experiment validates the crucial role and effectiveness of our pooling and resampler in extracting useful, high-quality information for the retrieval tasks.

We then investigate the components within our diversity loss, $\mathcal{L}_{\text{div}}$ (Eq. (5)). We first remove the Dropout mechanism

Table 12. **MMEBv2 Details with VLM2VecV2 and OmniRet**. **Bold** indicates the best performance in each dataset while underline indices the performance from training on the datasets.

| Task | Dataset | VLM2VecV2 [50] | OmniRet |
|---|---|---|---|
| I-CLS | ImageNet-1K | **80.8** | 57.0 |
| | N24News | **72.9** | 40.6 |
| | HatefulMemes | **56.3** | 49.9 |
| | VOC2007 | **85.0** | 68.1 |
| | SUN397 | **71.0** | 68.7 |
| | Place365 | 35.9 | **38.3** |
| | ImageNet-A | **47.4** | 39.1 |
| | ImageNet-R | **89.3** | 79.0 |
| | ObjectNet | **65.2** | 63.6 |
| | Country211 | **25.2** | 12.8 |
| I-RET | VisDial | **82.7** | 49.2 |
| | CIRR | **57.5** | 52.4 |
| | VisualNews_t2i | **74.5** | 73.5 |
| | VisualNews_i2t | **78.2** | 76.5 |
| | MSCOCO_t2i | **75.3** | 68.2 |
| | MSCOCO_i2t | **71.4** | 68.7 |
| | NIGHTS | **68.6** | 59.3 |
| | WebQA | **90.6** | 84.9 |
| | FashionIQ | 19.5 | **35.9** |
| | Wiki-SS-NQ | **66.9** | 41.9 |
| | OVEN | 64.3 | **80.1** |
| | EDIS | 84.1 | **92.5** |
| V-CLS | K700 | 38.0 | **45.3** |
| | SmthSmthV2 | 42.8 | **52.3** |
| | HMDB51 | 40.9 | **48.5** |
| | UCF101 | 60.0 | **65.7** |
| | Breakfast | 14.8 | **31.4** |
| V-RET | DiDeMo | 30.4 | **35.8** |
| | MSR-VTT | 28.3 | **38.7** |
| | MSVD | 48.1 | **64.9** |
| | VATEX | 26.5 | **33.6** |
| | YouCook2 | **10.6** | 9.6 |
| V-MRET | QVHighlight | 49.4 | **66.9** |
| | Charades-STA | 20.2 | **22.4** |
| | MomentSeeker | **40.8** | 40.5 |

from the loss function calculation. This resulted in a performance decrease of 1.5 score points. This result suggests that the unregularized diversity loss is overly aggressive when applied uniformly to all token pairs, indicating that some tokens must retain partial similarity to others. Additional experiments were conducted to determine the optimal regression target, comparing $\text{smooth}_{L1}$ against standard $L1$ and $L2$ norms. The use of $\text{smooth}_{L1}$ proved optimal, leading to an increase in recall of over 1.5 points, confirming its suitability for balancing distance and robustness in the diversity objective.

To understand the sensitivity of the Shared Media Resampler to feature dimensionality, we analyze the effect of decreasing the number of latents $N$. Instead of the optimal $N = 64$ tokens, we tested reduced sizes of $N = 32$ and $N = 16$. The corresponding performance dropped significantly by 2.1 and 2.5 score points, respectively. This clear correlation demonstrates that employing a larger number of media tokens is essential for successfully capturing the fine-grained information required for high-performance multimodal tasks.

Table 13. **Additional Ablation Studies on Proposed Modules.** We report the impact on Average Recall (across 6 tasks, trained on 1M samples) when each component is removed or modified.

| | Avg. Recall | Δ |
|---|---|---|
| Batch size = 1024 | | |
| **Our Baseline** | 49.7 | 0.0 |
| Multi (16) Vectors | 49.8 | +0.1 |
| No Resampler | 46.7 | -3.0 |
| Batch size = 2048 | | |
| **Our Baseline** | 50.2 | 0.0 |
| W/o Dropout in $\mathcal{L}_{\text{div}}$ | 48.7 | -1.5 |
| Replace $\text{smooth}_{L1}$ with L1 in $\mathcal{L}_{\text{div}}$ | 48.0 | -2.2 |
| Replace $\text{smooth}_{L1}$ with L2 in $\mathcal{L}_{\text{div}}$ | 48.5 | -1.7 |
| Sampled media tokens: $N = 32$ | 48.1 | -2.1 |
| Sampled media tokens: $N = 16$ | 47.7 | -2.5 |

Table 14. **Speed and Memory Usage of Retrieval Models on MSCOCO T2I.** The performance is measured on RTXA6000 with SDPA attention.

| Model | Speed (samples/s) | Mem (GB) |
|---|---|---|
| CLIP | 114.35 | 1.62 |
| SigLIP | 53.43 | 3.36 |
| PE-Core | 79.05 | 2.54 |
| $\text{CLIP}_{SF}$ | 116.04 | 1.62 |
| VLM2VecV2 | 10.21 | 4.39 |
| OmniRet | 35.69 | 6.01 |

## 9.6. Computational Cost

Table 14 presents the speed and memory usage of various retrieval models evaluated on the MSCOCO Text-to-Image (T→I) evaluation set. All models were benchmarked on the same NVIDIA RTX A6000 GPU utilizing SDPA (Scaled Dot-Product Attention) for consistent measurement. As expected, dual-encoder architectures, such as CLIP and SigLIP, demonstrate superior inference speed and significantly lower memory utilization. While VLM2Vec-V2 uses less memory than our OmniRet, our model achieves a substantial encoding speedup of 3.5× compared to it. The larger memory footprint observed in OmniRet is primarily attributed to the Perceiver blocks used to process the media tokens and the LLM hidden states.

Table 15. **Details of Table 2 with OmniRet's Stage 1 included**. Modalities: I (Image), V (Video), A (Audio), T (Text). **Bold** and **green** indicate the best performance in each group and overall, respectively. MMEmbed [40] is included for reference only, as it uses a larger LLM and is not directly comparable.

| Task | Dataset | Text-Binding Pretrained Models | | | | | Multi-Task Finetuning Models | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CLIP [57] | SigLIP [78] | PE-Core [3] | MMT [54] | CLAP [74] | CLIP$_{SF}$ [71] | MMEmbed [40] | VLM2VecV2 [50] | OmniRet (stage 1) | OmniRet (stage 2) |
| *Uni-Modality Tasks* | | | | | | | | | | | |
| I→I | NIGHTS | 25.9 | 25.9 | **32.0** | - | - | **28.4** | 32.1 | 30.0 | 27.3 | 25.0 |
| T→T | WebQA | 40.5 | 34.0 | **56.8** | - | - | 83.7 | 96.7 | 81.1 | **91.7** | 87.8 |
| *Text-Binding Tasks* | | | | | | | | | | | |
| I→T | VisualNews | 42.0 | 43.2 | **50.8** | - | - | **41.4** | 40.4 | 28.7 | 23.5 | 30.3 |
| | Fashion200K | 7.7 | **36.1** | 31.9 | - | - | 18.1 | 18.4 | 12.1 | 28.5 | **29.9** |
| | MSCOCO | 79.7 | 90.2 | **91.3** | - | - | **90.6** | 90.9 | 89.4 | 84.7 | 87.8 |
| T→I | VisualNews | 44.4 | 42.8 | **48.2** | - | - | **40.7** | 35.4 | 26.9 | 26.3 | 30.2 |
| | Fashion200K | 7.2 | **36.4** | 32.8 | - | - | 15.2 | 16.5 | 13.2 | 28.9 | **30.9** |
| | MSCOCO | 61.0 | 77.0 | **79.4** | - | - | **79.3** | 80.7 | 79.3 | 74.4 | 76.8 |
| V→T | TGIF | 20.6 | 25.1 | **35.8** | - | - | - | - | 17.3 | 14.7 | **33.8** |
| | Charades | 7.9 | 14.3 | **28.8** | - | - | - | - | 17.8 | 9.9 | **52.0** |
| T→V | TGIF | 19.8 | 27.5 | **34.2** | - | - | - | - | 18.2 | 14.4 | **33.4** |
| | Charades | 11.9 | 17.6 | 24.7 | - | - | - | - | 18.6 | 10.7 | **50.7** |
| A→T | AudioCaps | - | - | - | **76.8** | 76.7 | - | - | - | 76.9 | **81.9** |
| | ClothoV2.1 | - | - | - | 22.7 | **51.1** | - | - | - | **52.8** | 52.6 |
| T→A | AudioCaps | - | - | - | **72.0** | 70.3 | - | - | - | 73.8 | **76.5** |
| | ClothoV2.1 | - | - | - | 21.6 | **42.9** | - | - | - | 50.1 | **50.4** |
| *Composed Visual-Text Tasks* | | | | | | | | | | | |
| T→I,T | WebQA | - | - | - | - | - | 75.9 | 85.2 | 77.4 | 52.1 | **80.8** |
| | EDIS | - | - | - | - | - | 51.2 | 69.1 | 45.8 | 50.5 | 58.6 |
| I,T→T | OVEN | - | - | - | - | - | **48.8** | 46.6 | 28.4 | 15.8 | 45.9 |
| | INFOSEEK | - | - | - | - | - | 33.2 | 47.9 | 20.6 | 22.0 | **40.7** |
| I,T→I | FashionIQ | - | - | - | - | - | 16.5 | 25.5 | 15.2 | 3.4 | **27.3** |
| | CIRR | - | - | - | - | - | 36.3 | 51.5 | 42.1 | 12.1 | **46.8** |
| I,T→I,T | OVEN | - | - | - | - | - | **69.9** | 71.2 | 44.6 | 32.8 | 69.5 |
| | INFOSEEK | - | - | - | - | - | 51.4 | 65.9 | 22.5 | 25.5 | **58.1** |
| V,T→V | WebCoVR | - | - | - | - | - | - | - | 76.4 | 66.3 | **85.7** |