IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 March 2025, accepted 8 April 2025, date of publication 10 April 2025, date of current version 21 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3559663

RESEARCH ARTICLE

Synergy-CLIP: Extending CLIP With Multi-Modal Integration for Robust Representation Learning

SANGYEON CHO, (Graduate Student Member, IEEE), JANGYEONG JEON, (Graduate Student Member, IEEE), MINGI KIM, AND JUNYEONG KIM, (Member, IEEE)

Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea

Corresponding author: Junyeong Kim (junyeongkim@cau.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by Korean Government (MSIT) (No. RS-2023-00253782 and the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korean Government (MSIT) (No. 2021-0-01341) Artificial Intelligence Graduate School Program (Chung-Ang University).

ABSTRACT Multi-modal representation learning has become a pivotal area in artificial intelligence, enabling the integration of diverse modalities such as vision, text, and audio to solve complex problems. However, existing approaches predominantly focus on bimodal interactions, such as image-text pairs, which limits their ability to fully exploit the richness of multi-modal data. Furthermore, the integration of modalities in equal-scale environments remains underexplored due to the challenges of constructing large-scale, balanced datasets. In this study, we propose Synergy-CLIP, a novel framework that extends the contrastive language-image pre-training (CLIP) architecture to enhance multi-modal representation learning by integrating visual, textual, and audio modalities. Unlike existing methods that focus on adapting individual modalities to vanilla-CLIP, Synergy-CLIP aligns and captures latent information across three modalities equally. To address the high cost of constructing large-scale multi-modal datasets, we introduce VGG-sound+, a triple-modal dataset designed to provide equal-scale representation of visual, textual, and audio data. Synergy-CLIP is validated on various downstream tasks, including zero-shot classification, where it outperforms existing baselines. Additionally, we introduce a missing modality reconstruction task, demonstrating Synergy-CLIP's ability to extract synergy among modalities in realistic application scenarios. These contributions provide a robust foundation for advancing multi-modal representation learning and exploring new research directions. The code is available at https://github.com/JoSangYeon/Synergy-CLIP

INDEX TERMS Multi-modal, multi-modal representation learning, missing modality, missing modality reconstruction, speech and multi-modality, vision and language.

I. INTRODUCTION

Recent progress in the domain of artificial intelligence (AI) has revealed multi-modal learning as an innovative paradigm [1], [2], [3] that utilizes information spread over various interfaces, such as visual and auditory modalities. Multi-modal AI leverages diverse input data (i.e., multi-modal data) to solve complex problems, thereby overcoming the inherent limitations of models trained on single-modality data and exhibiting superior performance [4], [5]. Recent ongoing studies [6], [7] have focused on the fusion and utilization of various modalities, including discussions on the effectiveness and efficiency of modeling methodologies, to facilitate learning and improve them in deep learning applications such as computer vision [8], [9] and natural language processing [5], [10]. In addition, a groundbreaking approach known as contrastive language-image pre-training (CLIP) [11] has been introduced. CLIP utilizes large datasets for training by combining images and text, and efficiently links vision and language models. It aligns the representations in the two modalities while matching images to relevant text descriptions, focusing on understanding the interactions

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

between the input modalities. The innovative approach and robust performance of CLIP have contributed significantly to the advancement of multi-modal representation learning, marking an important milestone in subsequent research [12], [13], [14].

However, follow-up research on the CLIP model remains largely limited to exploring bimodal interactions [13], [15], [16], [17], such as image-text [11] and audio-text data [12]. The addition of new modalities primarily takes the form of adapting individual modalities within a bimodal framework [14], [18], [19], making it challenging to pursue research that learns and represents all modalities on an equal footing. These limitations stem from the high cost [20], [21] of annotating and acquiring large-scale tri-modal datasets for pre-training. However, to endow models with the ability to process and integrate diverse information like humans, there is a pressing need for research that encompasses multiple modalities with equal representation. Therefore, we underscore the importance of moving beyond bimodal approaches to emphasize the need for datasets that support tri-modal representation learning, facilitating the development of models that can process and understand complex information in a human-like manner. We envision this expansion will enable models to absorb and integrate diverse types of information, thus approaching human-like cognitive capabilities.

Motivated by the these limitations and the need for studies to process and integrate diverse information like humans, we introduce Synergy-CLIP, a novel framework for tri-modal representation learning. The objectives of Synergy-CLIP are illustrated in figure 1. It aligns and captures the latent information embedded within intricately intertwined tri-modal data. In addition, we propose the VGG-sound$^+$ dataset to pre-train Synergy-CLIP. The VGG-sound$^+$ dataset is a tri-modal dataset built upon VGG-sound [22], containing matched visual-audio-text data. It includes image-text-audio modalities by augmenting the visual-audio pairs with appropriate textual descriptions, resulting in a balanced dataset where all three modalities are represented at equal scale. Further, we utilize specific modules of the trained Synergy-CLIP to perform various validations. We perform the Missing Modality Reconstruction (MMR) task using learned high-dimensional representations to evaluate the effectiveness of Synergy-CLIP in capturing synergies between modalities and reconstructing missing modalities in various scenarios. Additionally, we validate the representational power of Synergy-CLIP using modality-specific downstream tasks, including zero-shot classification. The experimental results confirm that the proposed methodology outperforms the baseline in certain tasks.

This study makes the following contributions to multi-modal learning:

- **Introduction and explanation of the tri-modal dataset:** This study introduces a comprehensive tri-modal dataset, VGG-sound$^+$, which is expected to serve as a valuable resource for future research on tri-modal learning scenarios.
- **Proposal of Synergy-CLIP:** We propose Synergy-CLIP, a new model that extends the traditional CLIP framework to account for and process tri-modal data, creating a foundation for research to leverage diverse information in a human-like manner.
- **Introduction of a new validation methodology:** To validate the effectiveness of Synergy-CLIP, we introduce the MMR task. By evaluating the model's ability to reconstruct missing modal information, a more realistic validation metric is introduced for performance evaluation in real-life applications, engendering new perspectives on multi-modal learning.

This research not only paves the way for more sophisticated integration of multi modalities within AI systems, but also highlights the potential for leveraging different data types to understand and solve complex problems. Additionally, by leveraging MMR tasks, it will be possible to develop a wide range of applications in real-world domains, such as healthcare and natural sciences.

The remainder of this paper is organized as follows. Section II reviews related work on multi-modal representation learning and missing modality reconstruction. Section III introduces the Synergy-CLIP framework and the VGG-sound+ dataset. Section IV presents experimental results, highlighting Synergy-CLIP's performance. Section V concludes with contributions and future directions, and Section VI discusses limitations and ethics.

II. RELATED WORK

A. MULTI-MODAL REPRESENTATION LEARNING

Several studies have been conducted to understand and integrate multi-modal data within deep learning frameworks [23], [24]. These studies have made significant contributions in the field of multi-modal representation learning, which seeks to build integrated representations from various data sources, such as images, text, and audio data, or images from RGB-Depth-Segment [25], [26] modalities. Existing studies have primarily utilized transformer-based architectures [27], [28] to capture complex relationships based on inter-modality attention, exhibiting exceptional predictive power. Recently, CLIP has been introduced as a groundbreaking approach [11], which uses separate encoders for images and text based on contrastive learning to extract representations for each modality and align them. CLIP exhibits effective representation extraction performance by leveraging aligned representations, enabling zero-shot tasks to be performed based on integrated expressions. The ability of CLIP to utilize large datasets for training not only bridges the semantic gap between visual and textual data, but also unveils new directions for subsequent research in this field [12], [13], [14], [29], [30].

Building on the foundation established by CLIP, recent research has demonstrated the potential of broadening the

horizons of multi-modal representation learning comprehensively by exploring combinations of modalities beyond traditional image-text pairs, such as audio-text [12], [14] and 3D image-text [13] pairs. In this context, we aim to extend existing methodologies by simultaneously connecting not just two but three modalities. Previous research on expanding from bimodal to tri-modal analysis has focused on ''adapting'' [18], [19] representations of newly added modalities using the robustly trained vanilla-CLIP. However, this does not constitute an equitable expansion of the modality scale. This study emphasizes the necessity and importance of a dataset that represents the three most widely used data modalities, image, text, and audio, equally. Thus, our goal is to process and learn data from these three modalities on an equal footing to achieve high-dimensional representation learning that integrates and understands the modalities more comprehensively. Further, we investigate the impact of high-level integrated representations corresponding to various subtasks and modality combinations. The proposed Synergy-CLIP framework and our investigation enhances the understanding and integration of multi-modal data and suggests new directions for future research in the field of AI.

B. MISSING MODALITY RECONSTRUCTION

Recently, considerable research has been conducted to address the problem of missing modalities in multi-modal data [31], [32]. Previous studies have either advocated for learning shared features [33] across modalities or used Transformer Encoders to understand the relationships between missing modalities based on attention matrices [34], reporting improvements in robustness. In addition, methods such as Semi-Supervised Learning and Knowledge Distillation [35] have been used to learn representations of missing modalities from teacher models. Most previous studies have aimed to bridge the information gap induced by missing modalities during training or inference stages [34], [35]. Although effective, these approaches result in unconventional and complex architectures [33]. In contrast, we propose an MMR task that aims to emulate human cognitive abilities by reconstructing missing modalities using available ones. This task focuses on reconstruction, which differs significantly from existing methodologies in terms of its objectives for handling missing modalities. We utilize high-dimensional integrated representations constructed using the Synergy-CLIP architecture to reconstruct the features of the missing modality, and ultimately, the missing modality itself. To achieve this goal, we introduce a transformer-based encoder-decoder framework. The structure of the proposed model is illustrated in figure 3.

We consider three reconstruction scenarios based on the constructed Triplet Dataset (image-text-audio).

1) **Missing Image Reconstruction.** When the vision modality is missing, the proposed model uses only text and audio modalities to reconstruct the missing visual content. This approach demonstrates the model's capability to depict objects described by text-audio information visually.
2) **Missing Text Reconstruction.** When the text modality is missing, the proposed model uses the image and audio modalities to reconstruct a description or label that represents the provided content and context accurately. This showcases the model's ability to articulate what it ''sees'' and ''hears'' clearly.
3) **Missing Audio Reconstruction.** When the audio modality is missing, the proposed model uses the image and text modalities to reconstruct the missing auditory content. This demonstrates the model's ability to express auditory information clearly based on the provided visual content and descriptions.

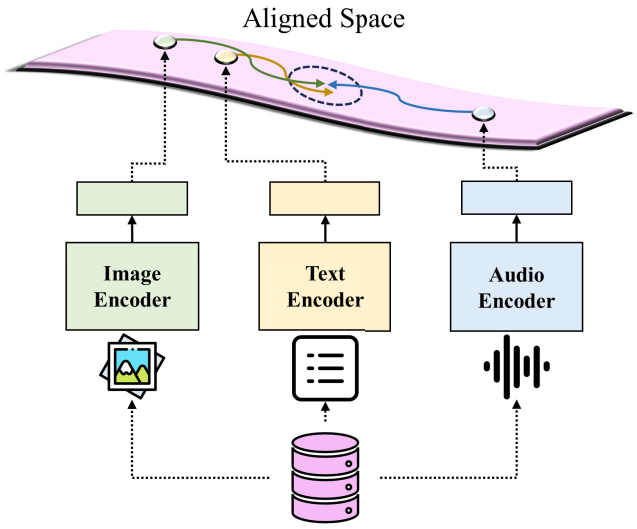


FIGURE 1. Mapping to aligned space for cross-modality. The aligned space acts as a convergence point for correlating different modal inputs and leveraging them for further multi-modal processing.

By proposing an approach for the reconstruction of multi-modal data, we introduce a new perspective on existing research directions and provide a foundation for the ultimate goals of AI technology, that is, emulating human cognitive abilities. Moreover, this design enhances the robustness of models concerning incomplete data and expands their applicability to scenarios in which data are inherently partial or subject to various limitations.

III. SYNERGY-CLIP

A. VGG-SOUND$^+$ BENCHMARK CONSTRUCTION

We construct the triple-modality dataset, VGG-sound$^+$, comprising image-text-audio data to pre-train Synergy-CLIP. VGG-sound [22] is a large-scale dataset designed for audio-visual learning, encompassing a diverse range of environmental and action sounds across various categories. It is widely utilized in the multimedia (audio-visual) domain and comprises over 310 distinct sound classes, including musical instruments, animal sounds, and mechanical noises. Each sample consists of an audio-visual pair extracted from YouTube videos, making it a valuable resource for

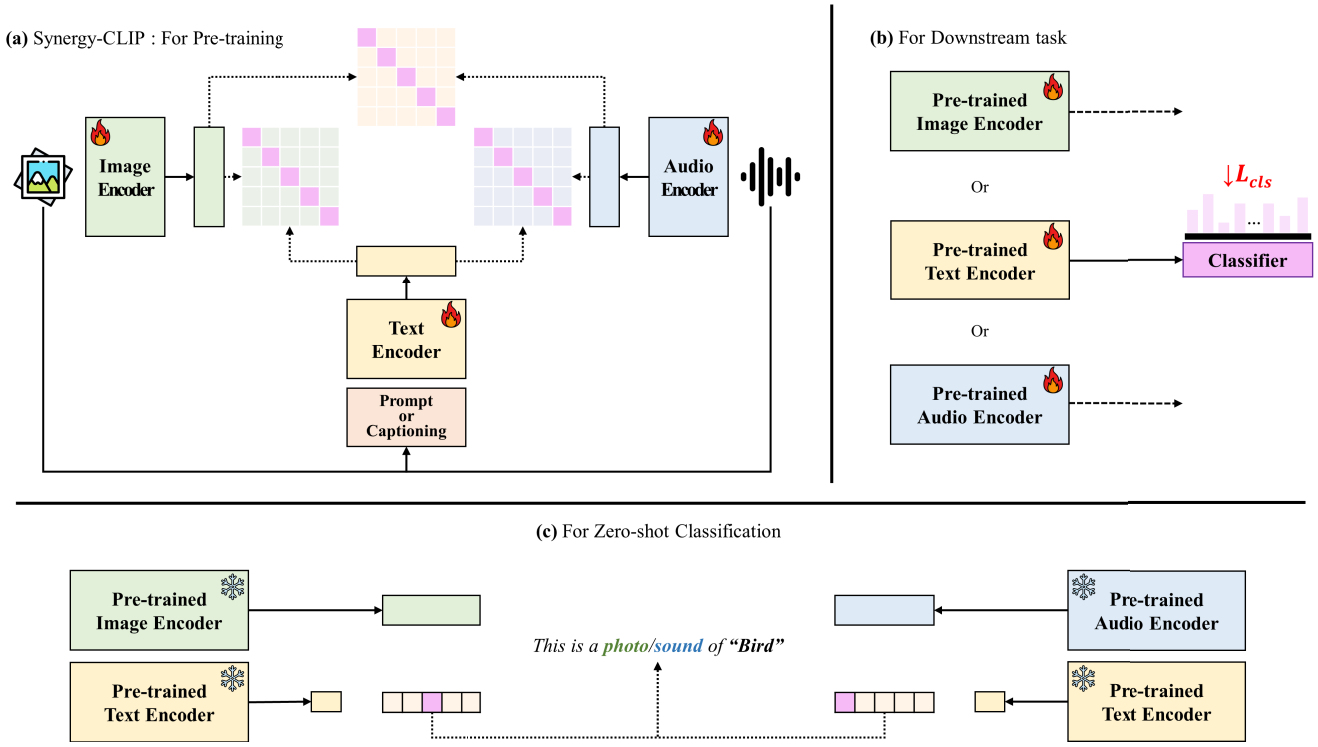


FIGURE 2. Illustration of Synergy-CLIP framework. (a) Pre-training of Synergy-CLIP involves inputs from three modalities: image, text, and audio. During pre-training, representations extracted by each modality-specific encoder are aligned, utilizing contrastive loss to facilitate this alignment. The pre-trained encoders of Synergy-CLIP are subsequently fine-tuned for modality-specific downstream tasks (b) as well as zero-shot classification (c), enabling the framework to adaptively leverage aligned representations across various AI tasks.

multi-modal learning and retrieval tasks. Based on VGG-sound, VGG-sound$^+$ consists of *200,000* audio-visual data entries categorized as video data, including metadata labeling the category of each video clip. We define the image-text-audio triplet modalities of VGG-sound$^+$ as the dataset $D_i = (I_i, T_i, A_i)$, where I_i represents an image snapshot of the video, T_i denotes a textual description of the video, and A_i signifies the audio clip. D_i is utilized during pre-training and on MMR tasks.

1) ADDING A TEXT DESCRIPTION

We extract random scenes from approximately 10 seconds-long video clips for I_i and capture the audio of the video clips for A_i. However, the VGG-sound dataset contains only metadata specifying the category of each video clip, without accompanying textual captions describing them in detail. Thus, we generate suitable captions T_i for the acquired I_i and A_i. Two approaches to captioning are considered: the semi-handcrafted method and the captioning model method. Inspired by previous image-text pre-training studies [11], the semi-handcrafted approach constructs text description sets by randomly combining categories recorded in metadata with appropriate prompt templates (e.g., *'a photo and sound of [category]'*). We devise 71 prompts for this training method. In contrast, the captioning model method uses extracted images and audio as inputs to generate text descriptions that include words recorded in the metadata. We use BLIP-2 [36] as the captioning model. The equation of the text description method, TD(·), for each T_i is as follows:

$$T_i = \begin{cases} \text{TD}_P(\text{template, category}) & \text{if method is } P \\ \text{TD}_C(I_i, A_i, \text{metadata}) & \text{if method is } C \end{cases} \quad (1)$$

where P denotes the semi-handcrafted approach and C represents the captioning model method. We anticipate that text descriptions generated using the captioning model provide the model with richer expressiveness.

B. TRI-MODAL REPRESENTATION LEARNING

We conduct pre-training using the constructed image-text-audio triplets to align the representations of each modality. In particular, we implement a model structure [37], [38], [39] that performs well corresponding to each modality and use it as a representation encoder for each modality. We define the representations to be aligned for each modality as follows:

$$\begin{aligned} h_i^{img} &= \mathsf{F}_{img}(I_i) \\ h_i^{txt} &= Avg_pool(\mathsf{F}_{txt}(T_i)) \\ h_i^{aud} &= \mathsf{F}_{aud}(A_i) \end{aligned} \quad (2)$$

where $\mathsf{F}_{img}(\cdot)$, $\mathsf{F}_{txt}(\cdot)$, and $\mathsf{F}_{aud}(\cdot)$ denote the encoders for the different modalities, and h_i^{img}, h_i^{txt}, and h_i^{aud} denote the respective modality representations extracted by these encoders.

1) ALIGNING REPRESENTATIONS OF TRIPLET MODALITIES

For each object i, I_i, T_i, and A_i denote the inputs into their respective encoders used to extract the representations, h_i^{img}, h_i^{txt}, and h_i^{aud}, respectively. These are aligned using the following equation:

$$\begin{aligned}\mathcal{L}_{total} = &\alpha \mathcal{L}_{clip}(h^{img}, h^{txt}) \\ &+ \beta \mathcal{L}_{clip}(h^{txt}, h^{aud}) \\ &+ \gamma \mathcal{L}_{clip}(h^{aud}, h^{img})\end{aligned} \tag{3}$$

where α, β, and γ denote hyper-parameters used to align the modality representations, with default values set to 1 for maintain balance among modality combinations. If α and β are 1 and γ is 0.5, the audio-image modality would be learned less effectively. Synergy-CLIP aims to capture interactions among modalities under equal conditions. Testing to prevent image-text over-fitting with α at 0.5 and β, γ at 1 showed inferior results, supporting the hypothesis for equal settings. A detailed analysis of the experimental results is provided in the Ablation Study section. Further, $\mathcal{L}_{clip}(\cdot)$ denotes the contrastive loss for paired modality representations. The detailed equation is as follows:

$$\begin{aligned}\mathcal{L}_{clip}(h^{M_1}, h^{M_2}) = -\frac{1}{N}\sum_{i=1}^{N}\Bigg[&\log \frac{\exp\left(\frac{sim(h_i^{M_1}, h_i^{M_2})}{\tau}\right)}{\sum_j^N \exp\left(\frac{sim(h_i^{M_1}, h_j^{M_2})}{\tau}\right)} \\ &+ \log \frac{\exp\left(\frac{sim(h_i^{M_1}, h_i^{M_2})}{\tau}\right)}{\sum_j^N \exp\left(\frac{sim(h_j^{M_1}, h_i^{M_2})}{\tau}\right)}\Bigg]\end{aligned} \tag{4}$$

where M_1 and M_2 are paired modalities, and h represents the modality vectors. The function $sim(\cdot)$ computes the similarity between two representations, and τ denotes a temperature parameter used to scale the similarity scores. $\mathcal{L}_{clip}$ maximizes the cosine similarity between correct modality pairs and minimizes it between incorrect pairs, thereby learning semantic relationships between the modalities.

C. MISSING MODALITY RECONSTRUCTION

The primary aim of AI development is to emulate human cognitive capabilities to develop versatile and powerful systems [40]. One of the remarkable abilities of humans is to perceive and reconstruct missing modalities using available sensory data. For instance, hearing a bird chirping might inspire the composition of a beautiful poetic verse or conjure a vivid visualization of the bird itself. To emulate such connections, we define and perform MMR, enabling AI systems to infer and reconstruct absent modalities. Although MMR is conceptually straightforward, its realization faces a complex challenge. Given a set of modalities, the goal is to reconstruct a missing modality using the information provided by the available modalities. By learning to infer and reconstruct absent forms, the proposed model enhances its resilience to data incompleteness, thereby improving its robustness and adaptability in real-world applications where data incompleteness or corruption are common.

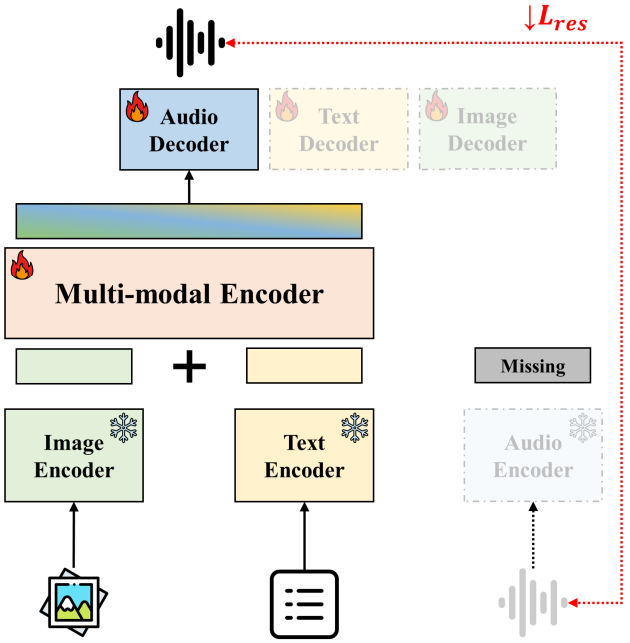


FIGURE 3. Framework for the missing modality reconstruction (MMR) task, this figure illustrates the use of a multi-modal encoder and a missing modality decoder to reconstruct the audio modality using representations extracted by Synergy-CLIP.

The proposed model structure for performing MMR is illustrated in figure 3. As illustrated in the figure, we design a multi-modal encoder and a missing modal decoder to execute MMR. For each modality, that is, image, text, and audio, representations are initially obtained using the pre-trained encoders. Subsequently, one modality input is masked, and the model is used to encode information about the missing modality using the remaining ones. For instance, if the audio modality is missing (see figure 3), the model relies on the image and text representations to encode information about the missing audio modality. The representation of the missing modality is then transmitted to the decoder, which reconstructs the modality. The model is trained by minimizing the loss $\mathcal{L}_{res}$ between the original and reconstructed forms.

1) OBJECTIVE FUNCTION FOR MMR

To optimize MMR, we define the following objective function:

$$\mathcal{L}_{\text{res}}(M, \hat{M}) = \begin{cases} \mathcal{L}_{res}^{img}(M, \hat{M}, \delta), & \text{if } M \text{ is } I \\ \mathcal{L}_{res}^{txt}(M, \hat{M}, \eta), & \text{if } M \text{ is } T \\ \mathcal{L}_{res}^{aud}(M, \hat{M}, \theta), & \text{if } M \text{ is } A \end{cases} \tag{5}$$

where $\mathcal{L}_{\text{res}}(M, \hat{M})$ is calculated dynamically for the image (I), text (T), and audio (A) modalities. Additionally, M represents the original modality and $\hat{M}$ represents the modality reconstructed by the model.

If the image (I) modality is missing, the objective function becomes $\mathcal{L}_{res}^{img}(I, \hat{I}, \delta)$. $\mathcal{L}_{res}(\cdot)$ is composed of SSIM and ℓ_2, where SSIM captures the essential structural information present in visual data, and ℓ_2 preserves pixel-level details. The complete expression for image reconstruction is as follows:

$$\mathcal{L}_{res}^{img}(I, \hat{I}, \delta) = \delta \cdot \text{SSIM}(I, \hat{I}) + (1 - \delta) \cdot \mathcal{L}_{\ell_2}(I, \hat{I}) \quad (6)$$

where the default value of δ is set to 0.75. This is because a greater weight needs to be assigned to SSIM during image reconstruction as it captures colorful details more effectively than ℓ_2-loss.

If the text (T) modality is missing, the *cross entropy loss* loss function is employed, which is effective in the context of categorical distributions in textual data. The loss function is expressed as follows:

$$\mathcal{L}_{res}^{txt}(M, \hat{M}, \eta) = \eta \cdot \text{CE-Loss}(T, \hat{T}) \quad (7)$$

where the default value of η is 1.0.

If the audio (A) modality is missing, the objective function is taken to be $\mathcal{L}_{res}^{aud}(A, \hat{A}, \theta)$. $\mathcal{L}_{res}^{aud}(\cdot)$ includes SSIM and ℓ_2. As in the case of images, SSIM captures the structural aspects of audio waveforms while ℓ_2-loss ensures the preservation of waveform pixel values. The complete expression for audio reconstruction is as follows:

$$\mathcal{L}_{res}^{aud}(A, \hat{A}, \theta) = \theta \cdot \text{SSIM}(A, \hat{A}) + (1 - \theta) \cdot \mathcal{L}_{\ell_2}(A, \hat{A}) \quad (8)$$

where the default value of θ is 0.25. In contrast to image reconstruction, audio reconstruction involves only a single channel. Therefore, θ is set to prioritize $\ell_2$loss, which excels at approximating the pixel values themselves.

During MMR training, the gradients are managed such that the weights of the modality-specific encoders, which are learned during the pre-training process, are frozen. Only the multi-modal encoder and the missing modal decoder are updated.

TABLE 1. Default pre-training setting.

Hyper-parameters	Value
Optimizer	AdamW
learning rate	5e-6
Weight decay	1e-5
Adam β	(0.9, 0.999)
Batch size	256
Training epochs	10
projection dim	1024
α in equation 3	1.0
β in equation 3	1.0
γ in equation 3	1.0
Image resolution	224 x 224
Image augmentation	RandomResizedCrop RandomHorizontalFlip RandomVerticalFlip
Text max length	32
Audio sample rate	16000
Audio Augmentation	AddWhiteNoise Shifting Stretching Flipping

TABLE 2. Default MMR task setting.

Hyper-parameters	Value
Optimizer	AdamW
learning rate	1e-4
Weight decay	1e-5
Adam β	(0.9, 0.999)
Batch size	128
Training epochs	200
projection dim	1024
δ in equation 5	0.75
η in equation 5	1.0
θ in equation 5	0.25
Image resolution	224 x 224
Text max length	32
Audio sample rate	16000

TABLE 3. Default fine-tuning & zero-shot setting.

Hyper-parameters	Value
Optimizer	AdamW
learning rate	1e-5
Weight decay	1e-5
Adam β	(0.9, 0.999)
Batch size	256
Training epochs	10
Image resolution	224 x 224
Image augmentation	RandomResizedCrop RandomHorizontalFlip RandomVerticalFlip
Text max length	32
Audio sample rate	16000
Audio Augmentation	AddWhiteNoise Shifting Stretching Flipping

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed Synergy-CLIP on various downstream tasks, including zero-shot classification, and assess the MMR solution performance of the multi-modal encoder and missing modal decoder frameworks based on Synergy-CLIP.

A. IMPLEMENTATION DETAILS

During pre-training, we implement the Synergy-CLIP$_{base}$ and Synergy-CLIP$_{large}$ models based on each backbone network (see figure 2 top-left). Additionally, text descriptions

TABLE 4. Performance of MMR. This table presents the reconstruction performance in various scenarios with missing modalities. Training reports on randomly sampled subsets of triplet datasets are listed, which contain similar meta-information at a rate of 10% of the total data.

Modality		Image			Text			Audio		
Captioning	Size	LPIPS(↓)	PSNR (↑)	SSIM (↑)	METEOR(↑)	FENSE(↑)	Accuracy(↑)	MCD(↓)	PSNR (↑)	SSIM (↑)
Prompt	Base	0.17	21.30	0.89	55.86	66.08	67.32	5.19	14.17	0.67
	Large	0.15	22.19	0.92	57.53	68.17	73.54	4.41	12.07	0.88
Caption	Base	0.17	21.32	0.89	57.32	69.36	75.46	**3.25**	**14.36**	0.87
	Large	**0.14**	**22.27**	**0.92**	**58.07**	**69.98**	**75.62**	3.38	14.34	**0.92**

are generated using semi-handcrafted prompts (P) and BLIP-2 [36] captions (C) during pre-training. We use four A6000 GPUs for training. For detailed pre-training settings, please refer to table 1. Reproducibility was ensured by fixing the Seed at 17.

For Missing Modality Reconstruction (MMR) task, a multi-modal encoder and missing modal decoder-based framework is devised based on pre-trained Synergy-CLIP to reconstruct the missing modalities. The encoder is designed to fuse representations of the remaining modalities. The decoders for image and audio uses a CNN-based structure, and the text decoder is based on a transformer-based structure. We also construct and train the models for various missing scenarios based on model size and text description (see figure 3). We use a pair of A6000 GPUs. For detailed MMR-training settings, please refer to table 2. Reproducibility was maintained with a fixed seed of 42.

In downstream tasks, we also perform experiments to validate the expressive power of pre-trained Synergy-CLIP. Common downstream tasks, primarily in the fields of vision, text, and audio, are considered for validation. Each modality-specific encoder is fine-tuned along with a classifier to perform tasks ranging from conventional classification to zero-shot classification (see figure 2 top right & bottom). For each task, we fixed Seeds at three values (17, 42, 77) to provide an average of performance and its standard deviation. For detailed pre-training settings, please refer to table 3.

B. BACKBONE NETWORKS

We configure the following backbone networks for each modality:

For the vision model, the ViT-base [37] and ViT-large architectures are chosen owing to their proven efficiency in capturing complex visual patterns using a self-attention mechanism, which considers task-related image areas flexibly.

For the text model, we use the RoBERTa-base [39] and RoBERTa-large architectures. Owing to more focused pre-training using dynamic masking, these models are effective in capturing complex text and integrated features over various modalities, and they outperform BERT [41].

For the audio model, we utilize AST [38], whose Base and Large models differ not in the number of parameters but in the overall patch size. AST is also a transformer-based architecture, which is capable of modeling and capturing complex acoustic patterns effectively.

C. DATASETS

For each modality, we evaluate the representations of pre-trained Synergy-CLIP on the following downstream tasks:

1) IMAGE DATASETS

Oxford-IIIT Pets [42], a dataset of images featuring 37 common breeds of pets, with approximately 200 images per breed, is primarily used to evaluate fine-grained visual classification algorithms. Flowers-102 [43], composed of images of 102 types of commonly seen flowers, with each category containing between 40 and 258 images, is used as a challenging dataset for fine-grained image recognition tasks. Subsets of a labeled portion from a collection of 80 million tiny images, CIFAR [44], CIFAR-10, containing 6,000 images for each of 10 classes, and CIFAR-100, containing 600 images for each of 100 classes, are used to evaluate image recognition algorithms.

2) TEXT DATASETS

GLUE Dataset [45], the General Language Understanding Evaluation (GLUE) benchmark, which aggregates various natural language understanding tasks, includes tasks for evaluating and analyzing model performance across various linguistic phenomena, such as sentiment analysis, question answering, and text implication.

3) AUDIO DATASETS

ESC-50 [46], a dataset composed of short, labeled environmental sound clips spread over 50 classes, serves as a standardized dataset for benchmarking models for environmental sound classification. It is also used to compare various methodologies in audio recognition tasks. We follow the dataset's recommendations for training/validation using 5-fold cross-validation. UrbanSound8k [47], a dataset containing 8,732 labeled short sound clips recorded in urban environments, is classified into ten classes. It is extensively used to evaluate the performance of urban sound classification algorithms, emphasizing the importance of audio signal processing in machine learning. We follow the dataset's recommendations for training/validation using 10-fold cross-validation.

D. MISSING MODALITY RECONSTRUCTION

In this section, we validate the performance of the proposed methodology in the MMR task both quantitatively and qualitatively.

TABLE 5. This table presents the performance of the image classification downstream tasks. Synergy-CLIP improves upon the performance of baselines within the same CLIP series on some tasks. The average values are reported along with the corresponding standard deviations over three different seed values.

	Oxford-IIIT Pets	Flowers-102	CIFAR-10	CIFAR-100	avg
ResNet 152 [11]	93.00	89.60	93.50	78.00	88.53
BiT-M [48]	92.40	99.30	97.60	88.20	94.38
ViT-B [37]	90.40	98.70	96.70	86.30	93.03
ViT-L [37]	92.90	99.30	97.90	89.00	94.78
CLIP : ViT-B [11]	90.00	96.90	95.10	80.50	90.63
CLIP : ViT-L [11]	95.10	99.20	98.00	87.50	94.95
Synergy-CLIP			Ours		
Base_{P} : vision enc	91.99 ± 0.05	97.68 ± 0.02	97.31 ± 0.01	83.28 ± 0.04	92.57 ± 0.03
Base_{C} : vision enc	91.76 ± 0.01	97.22 ± 0.04	97.10 ± 0.01	87.25 ± 0.03	93.33 ± 0.02
Large_{P} : vision enc	94.40 ± 0.02	**99.54 ± 0.01**	98.22 ± 0.01	**90.82 ± 0.02**	**95.74 ± 0.01**
Large_{C} : vision enc	**94.73 ± 0.01**	98.80 ± 0.08	**98.39 ± 0.02**	90.74 ± 0.02	95.67 ± 0.03

TABLE 6. This table presents the results of the GLUE Benchmark. In certain tasks, Synergy-CLIP exhibits superior performance compared to pre-trained models of similar size. The best performance out of the three seed-trained outcomes is reported here.

	MNLI-M	MNLI-MM	QNLI	QQP	SST-2	CoLA	MRPC	RTE	avg
BiLSTM+Attn, ELMo	72.40	72.40	75.20	83.60	91.50	44.10	82.10	52.70	71.75
Bert-base [41]	84.00	84.20	91.00	87.60	92.60	60.30	90.20	69.50	82.43
RoBERTa [39]	90.20	90.20	94.70	92.20	96.40	68.00	90.90	86.60	88.65
GPT-2 XL [49]	-	71.50	76.70	70.10	91.30	10.50	69.40	56.80	63.76
GPT-3 [50]	-	90.20	92.70	81.10	97.10	60.30	86.50	83.70	84.51
CLIP : txt enc. [29]	71.80	72.50	73.00	82.70	86.20	6.60	81.40	53.80	66.00
Synergy-CLIP				**Ours**					
Base_{P} : txt enc.	87.31	87.04	91.60	90.10	93.12	69.12	90.09	61.73	83.76
Base_{C} : txt enc.	87.53	87.31	91.93	90.11	93.35	69.13	89.30	65.70	84.30
Large_{P} : txt enc.	90.23	90.40	94.45	**91.74**	96.10	83.89	89.56	74.01	88.80
Large_{C} : txt enc.	**90.29**	**90.49**	**94.75**	91.71	**96.22**	**85.23**	**90.80**	**79.42**	**89.86**

TABLE 7. This table presents the results achieved in the sound classification downstream task. Synergy-CLIP outperforms the baseline in some tasks. Results are reported as the average value with the standard deviation over three seed values. The best performance compared to the baselines is marked with an asterisk(*).

	ESC-50	UrbanSound8k
Human	81.50	-
ESResNeXt [51]	95.20	89.14
AST [38]	95.60	-
AudioCLIP [14]	97.15	90.07
CLAP [12]	96.70	87.96
Synergy-CLIP	**Ours**	
Base_{P} : aud enc.	95.05 ± 0.02 (***97.75**)	**89.31 ± 0.03** (*94.44)
Base_{C} : aud enc.	94.55 ± 0.01 (*96.75)	89.18 ± 0.02 (***94.55**)
Large_{P} : aud enc.	94.70 ± 0.02 (*96.75)	87.83 ± 0.04 (*94.02)
Large_{C} : aud enc.	**95.10 ± 0.02** (*96.75)	88.50 ± 0.03 (*91.45)

1) QUANTITATIVE EVALUATIONS

Table 4 presents the quantitative evaluations of the encoder-decoder structure for the Multi-Modal Reconstruction (MMR) task, assessing the reconstruction performance across different missing-modality scenarios. For each modality, we employ appropriate evaluation metrics to measure the quality of the reconstructed outputs. The results highlight that models pre-trained with captioned data consistently outperform those trained with handcrafted prompts. Furthermore, increasing the model size (*base* → *large*) leads to a noticeable improvement in reconstruction performance across all modalities. For scenarios where the **image modality is missing**, we evaluate the reconstructed images using Peak Signal-to-Noise Ratio (PSNR) [52], Structural Similarity Index Measure (SSIM) [53], and Learned Perceptual Image Patch Similarity (LPIPS). LPIPS [54] measures perceptual similarity, where lower values indicate higher reconstruction quality. As shown in Table 4, larger models demonstrate superior reconstruction performance, achieving higher PSNR and SSIM scores while reducing LPIPS. The caption-based large model achieves the best image reconstruction results (PSNR: 22.27, SSIM: 0.92, LPIPS: 0.14), surpassing both prompt-based and base-sized models. In cases where the **text modality is missing**, we assess the reconstructed text using METEOR [55] and FENSE [56] scores, as well as accuracy, which measures the alignment between the reconstructed and ground-truth tokens. As observed in Table 4, models trained with captioned data exhibit stronger text reconstruction performance compared to those trained with handcrafted prompts. The caption-based large model achieves the highest scores (METEOR: 58.07,

TABLE 8. This table presents the zero-shot classification performance of Synergy-CLIP for image-text modality, with respect to the number of modalities aligned during pre-training. The data indicate that aligning the representations of three modalities yields better outcomes compared to aligning just two. Further, pre-training with captions instead of simple prompts yields enhanced representations.

Captioning Module	Size	Modality aligned	CIFAR-10		CIFAR-100	
			top - 1	top - 5	top - 1	top - 5
Prompt	Base	I + T	76.04 ± 0.04	99.74 ± 0.04	19.01 ± 0.04	41.15 ± 0.08
		I + T + A	76.56 ± 0.06	98.43 ± 0.01	22.01 ± 0.03	45.83 ± 0.03
	Large	I + T	84.04 ± 0.01	98.96 ± 0.03	24.23 ± 0.02	49.22 ± 0.03
		I + T + A	**84.38 ± 0.01**	**99.22 ± 0.01**	**28.26 ± 0.02**	**54.04 ± 0.02**
Caption	Base	I + T	69.53 ± 0.02	98.44 ± 0.01	16.92 ± 0.03	38.80 ± 0.02
		I + T + A	77.60 ± 0.02	97.66 ± 0.01	24.48 ± 0.03	50.13 ± 0.04
	Large	I + T	70.57 ± 0.03	98.18 ± 0.01	21.35 ± 0.02	50.26 ± 0.07
		I + T + A	**86.20 ± 0.05**	**99.22 ± 0.01**	**31.51 ± 0.03**	**61.07 ± 0.04**

TABLE 9. This table presents the zero-shot classification performance of Synergy-CLIP for audio-text modality. As with table 8, we see that aligning the three modalities performs better, and the Caption has a better representation.

Captioning Module	Size	Modality aligned	ESC-50		UrbanSound8k	
			top - 1	top - 5	top - 1	top - 5
Prompt	Base	A + T	56.25 ± 0.07	85.45 ± 0.02	48.56 ± 0.03	84.38 ± 0.01
		A + T + I	**66.88 ± 0.04**	88.59 ± 0.02	52.81 ± 0.06	**89.84 ± 0.02**
	Large	A + T	56.25 ± 0.02	86.72 ± 0.05	51.56 ± 0.05	84.22 ± 0.02
		A + T + I	65.47 ± 0.04	**88.75 ± 0.03**	**53.98 ± 0.05**	85.23 ± 0.06
Caption	Base	A + T	44.79 ± 0.05	76.82 ± 0.02	39.32 ± 0.03	88.80 ± 0.01
		A + T + I	59.84 ± 0.04	85.63 ± 0.03	46.48 ± 0.08	85.08 ± 0.03
	Large	A + T	55.83 ± 0.02	78.13 ± 0.01	51.30 ± 0.08	**92.96 ± 0.05**
		A + T + I	**66.25 ± 0.02**	**87.66 ± 0.03**	**58.59 ± 0.05**	90.00 ± 0.03

TABLE 10. Impact of different α, β, γ values on R@10 performance. The baseline configuration ($\alpha = \beta = \gamma = 1.0$) achieves the highest overall R@10 score (90.06%), indicating that balanced weight settings yield the best modality alignment. Reducing any of the weights individually (A1, A2, A3) leads to a drop in performance, especially in the text-audio and audio-image retrieval tasks. When multiple weights are reduced simultaneously (B1, B2, B3), the decline becomes more significant, confirming the necessity of equal weighting for optimal retrieval performance.

	α	β	γ	$\mathbf{R@10}^{img-txt}$	$\mathbf{R@10}^{txt-aud}$	$\mathbf{R@10}^{aud-img}$	$\mathbf{R@10}^{avg}$
Baseline	1.0	1.0	1.0	**96.25**	**89.65**	**84.29**	**90.06**
A1	0.5	1.0	1.0	94.17	87.23	83.17	88.19
A2	1.0	0.5	1.0	95.72	83.96	83.29	87.66
A3	1.0	1.0	0.5	94.79	87.44	81.12	87.78
B1	1.0	0.5	0.5	95.85	82.92	82.28	87.02
B2	0.5	1.0	0.5	94.27	88.62	81.98	88.29
B3	0.5	0.5	1.0	94.38	82.44	83.94	86.92

FENSE: 69.98, Accuracy: 75.62). For **missing audio modality** scenarios, we evaluate reconstruction quality using Mel Cepstral Distortion (MCD) [57] (where lower values indicate better quality) along with PSNR and SSIM. The results reveal that models pre-trained with captions outperform those trained with prompts, with the caption-based large model achieving the lowest MCD (3.38) and the highest PSNR (14.34) and SSIM (0.92). This demonstrates that caption-based training facilitates better alignment between audio and other modalities, leading to improved reconstruction quality.

2) QUALITATIVE EVALUATION

We also present a qualitative evaluation in figure 4. The figure 4 illustrates the inference results in MMR scenarios obtained using the best-performing Large-Caption model. Both image and audio results are observed to be highly similar to the actual outputs. The color differences likely result from the impact of hyper-parameters in equation 6 and 8 and the reconstruction performance of the decoders. Adjusting the boundaries of δ and θ in equation 6 and 8, which balance SSIM and L2 loss, is necessary. SSIM loss emphasizes brightness, contrast, and luminance, causing certain parts of

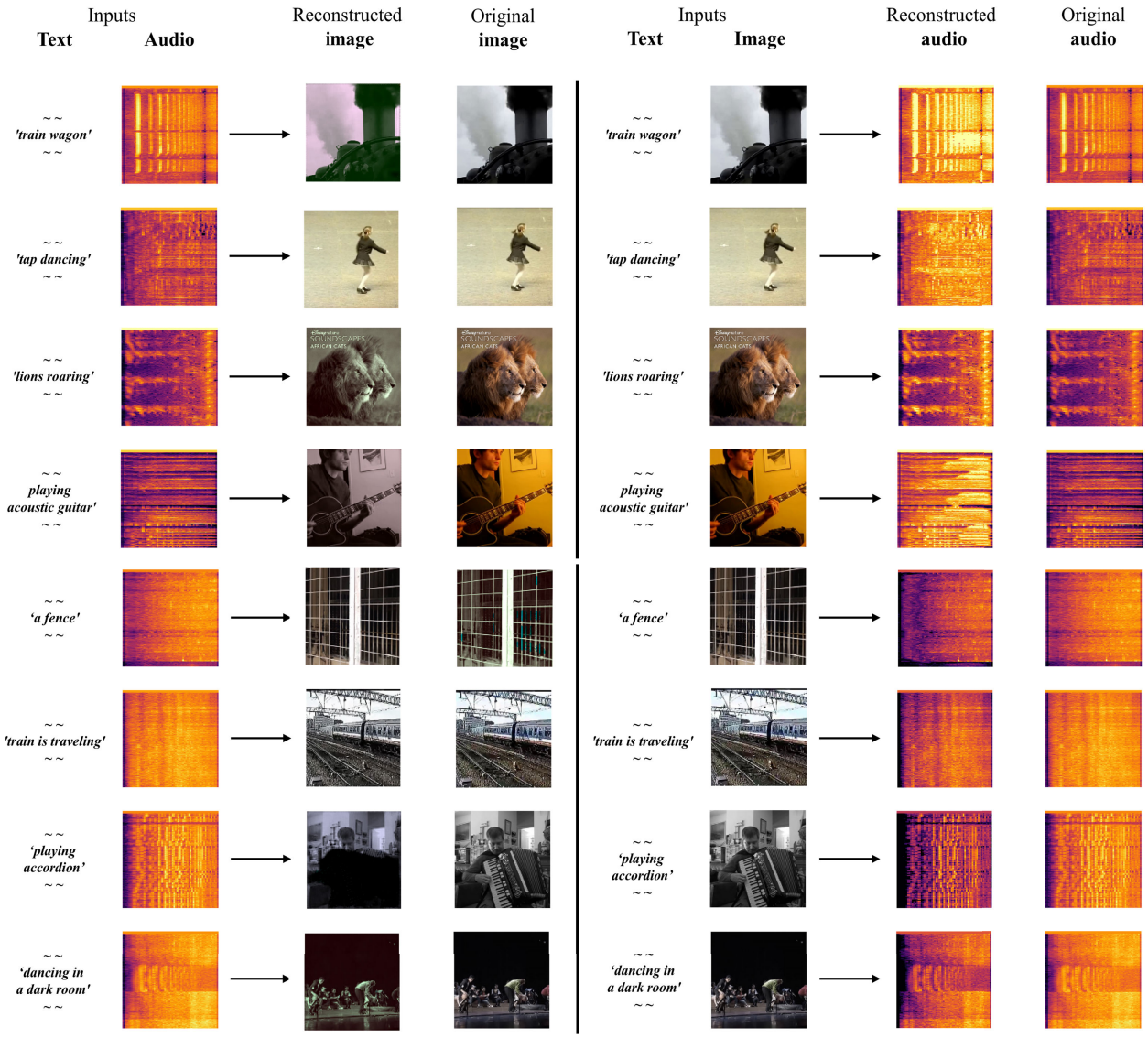


FIGURE 4. A qualitative example of the MMR task. In this evaluation, the text reconstruction scenario has been excluded (as it is a token classification). The figure illustrates the reconstruction results of the large caption model, which exhibits the best performance as per quantitative evaluations.

reconstructed modality to be highlighted differently. Also, We observed that the model produced outputs for both images and audio that closely resembled real examples across various categories. However, in certain samples, reconstruction failed, particularly with more complex objects such as musical instruments. This suggests that cross-modal representations in audio or text are weaker in these cases, highlighting a key challenge for MMR to address in future work. Nonetheless, the model demonstrates strong qualitative performance in reconstruction overall.

E. EXPERIMENTS BASED ON DOWNSTREAM TASKS

1) SINGLE-MODAL RESULTS

We perform and validate downstream tasks for each modality. The results for image, text, and audio data are presented in table 5, 6 and 7, respectively.

In Image Tasks: we found that Synergy-CLIP outperformed the baseline and a model trained in a similar context (CLIP [11] series). In particular, the Large models achieved the highest performance on all datasets. This means that Synergy-CLIP has learned a more sophisticated feature extraction process.

In Text Tasks: Moreover, the proposed model outperforms the baseline. In addition, it outperforms RoBERTa [39] on average, despite statistically insignificant differences on some tasks. This confirms that Synergy-CLIP has the ability to generalize across a variety of natural language understanding tasks.

In Audio Tasks: Finally, the proposed model outperforms similarly trained models (AudioCLIP [14] and CLAP [12]). These results underscore the model's ability to capture nuanced audio features, further supporting its effectiveness in extracting modality-specific representations. Overall, the

downstream task evaluations across the three modalities validate Synergy-CLIP's capability to enhance performance across diverse tasks, achieving state-of-the-art results or competitive performance. The model's ability to outperform specialized models in each modality—image, text, and audio—highlights its potential for multi-modal applications, where robust feature extraction and representation are crucial. The consistency of high performance across different datasets and tasks further substantiates Synergy-CLIP as an effective multi-modal learning framework.

2) ZERO-SHOT CLASSIFICATION RESULTS

Finally, we conduct and validate zero-shot classification for the previously performed downstream tasks. Zero-shot classification is highly influenced by the scale and quality of the pre-training data; therefore, to ensure fair evaluation, we report the results using models pre-trained in the same way as the previously introduced data. The performance of Synergy-CLIP in zero-shot classification is presented in table 8 and 9. The *"Modality aligned"* column in tables lists the results with respect to combinations of modalities with aligned representations during pre-training. As indicated in the table 8, 9, Synergy-CLIP, which aligns all three modalities, generally exhibits superior performance. Additionally, pre-training with captions, rather than simple prompts, yields enhanced expressiveness. These results confirm that the number of modalities aligned impacts the ability to capture latent relationships among them significantly and that using well-articulated captions over simple prompts enhances the expressive power of CLIP-series models.

F. ABLATION STUDY ON MODALITY ALIGNMENT

A controlled study was conducted to analyze the effects of the pre-training objective function defined in Eq.3. The experiments primarily focused on two aspects: the impact of reducing each hyperparameter individually and the effect of jointly decreasing two hyperparameters. After pre-training on VGG-sound$^+$, we evaluated the R@10 performance on the test subset of VGG-sound$^+$. The detailed results are provided in Table10.

In the experiments where one of α, β, or γ was individually adjusted (A1, A2, A3), a noticeable decline in R@10 was observed, highlighting the importance of balanced hyperparameter settings. Specifically, as the alignment strength between certain modalities weakened due to altered training parameters, the overall performance deteriorated accordingly.

Moreover, the experiments in which two of α, β, or γ were simultaneously reduced (B1, B2, B3) exhibited even more pronounced performance degradation. These results clearly demonstrate that weakening the alignment between specific modalities negatively impacts the alignment and performance of other modalities as well.

The findings indicate that maintaining equal weighting across all modalities yields the most optimal performance. When any of α, β, or γ was reduced, the R@10 performance dropped significantly, with the most substantial decline occurring in text-audio and audio-image retrieval. These results further substantiate the critical role of balanced modality alignment in ensuring robust multi-modal learning.

V. CONCLUSION AND FUTURE WORK

In this study, we propose Synergy-CLIP, an extension of the CLIP framework designed to enhance the expressiveness of models across visual, textual, and audio modalities and emulate human cognitive abilities. By extending the modalities and framework, we overcome the limitations of existing studies on paired modalities and achieve significant advancements. We also introduce the Missing Modality Reconstruction (MMR) task and perform comprehensive evaluations on various downstream tasks, including zero-shot classification. The results further demonstrate the robustness and flexibility of Synergy-CLIP.

This study emphasizes the importance of a holistic approach to multi-modal learning, in which coordinating and integrating multiple modalities can significantly improve model performance in various tasks. In particular, zero-shot classification results reveal that models pre-trained with more nuanced and comprehensive data (e.g., captions rather than simple prompts) exhibit superior performance, underscoring the importance of rich and descriptive pre-training in enhancing model functionality. Additionally, the MMR task, involving the reconstruction of missing sensory information based on available data, serves as a unique metric for assessing the resilience of models in real-world application environments. It not only tests the limits of current AI capabilities but also establishes a foundation for future research on understanding and interpreting incomplete or partial data.

The learning approach of Synergy-CLIP can be extended to incorporate a broader range of modalities, further enhancing the model's sensory integration capabilities. Additionally, by advancing the MMR task, the model can capture more nuanced interactions across modalities, thereby supporting various real-world applications through the reconstruction of missing information. Future research will focus on securing a richer diversity of modality data to reinforce inter-modality interactions robustly and on refining the MMR task to elevate its applicability in practical domains such as healthcare and security. Furthermore, the introduction of additional modalities will enable the exploration of a wider range of missing modality reconstruction scenarios, allowing Synergy-CLIP to handle more complex and diverse real-world cases. By addressing these challenges, future iterations of the model aim to further enhance its adaptability and effectiveness across multiple domains.

Overall, Synergy-CLIP sets a new benchmark for accommodating diverse modalities, potentially having a significant impact on multi-modal representational learning and serving as a pivotal foundation for understanding the complex interplay between various modalities.

VI. LIMITATION AND ETHICS

A. LIMITATIONS

While Synergy-CLIP demonstrates robust performance leveraging multi-modal datasets, it also presents limitations under specific conditions and assumptions. In terms of scalability, adding new modalities requires extensive resources for constructing and maintaining large-scale datasets that equally represent each modality. Furthermore, the additional loss term in equation 3 increases computational costs, potentially constraining practical scalability. Specifically, when introducing a new modality, the pre-training objective function incurs an ${}_nC_2$ computational overhead due to pairwise modality alignment, significantly increasing complexity as the number of modalities grows. Additionally, in the Missing Modality Reconstruction (MMR) task, Synergy-CLIP may struggle to reconstruct complex data, potentially impacting performance in applications that require precise restoration. Another limitation arises from BLIP-generated captions, which, while effective for multi-modal alignment, can be susceptible to inherent biases within the BLIP model. These biases may propagate into the training process, affecting the quality and diversity of textual representations. Addressing these biases is crucial to ensuring fair and unbiased learning across different modalities.

B. ETHICS

The ethical considerations surrounding Synergy-CLIP must be carefully addressed in both research and real-world applications. First, there is the issue of data bias. As Synergy-CLIP trains on large multi-modal datasets, it may inherit latent biases present within these datasets, which could reinforce or amplify societal biases in sensitive application areas. Continuous monitoring and dataset review during model training are essential to mitigate this risk. Additionally, the use of BLIP-generated captions introduces another layer of potential bias. Since BLIP itself is trained on large-scale vision-language datasets, it may exhibit biases in how it describes images, leading to skewed or stereotypical textual representations that can propagate through Synergy-CLIP's training pipeline. Careful curation and bias auditing of generated captions are necessary to prevent unintended reinforcement of these biases.

Second, there are privacy concerns. The MMR task, which infers missing visual or audio data based on existing information, could pose privacy risks in contexts where personal or public system data privacy is critical. In particular, the ability to reconstruct missing modalities raises concerns about potentially inferring sensitive or unintended information from incomplete data. Consequently, Synergy-CLIP applications, particularly those involving MMR, should operate under strict ethical oversight, ensuring that the model does not inadvertently infer sensitive or personally identifiable information.

REFERENCES

[1] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," 2021, *arXiv:2107.00135*.

[2] K. Zhou, C. Chen, B. Wang, M. R. U. Saputra, N. Trigoni, and A. Markham, "VMLoc: Variational fusion for learning-based multimodal camera localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, May 2021, pp. 6165–6173.

[3] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, Feb. 2023.

[4] A. Taleb, C. Lippert, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, Jan. 2021, pp. 661–673.

[5] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, May 2022.

[6] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (Provably)," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 10944–10956.

[7] S. Qu, Y. Kang, and J. Lee, Janghwan, "Efficient multi-modal fusion with diversity analysis," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2663–2670.

[8] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets," *Vis. Comput.*, vol. 38, no. 8, pp. 2939–2970, Aug. 2022.

[9] W. Chai and G. Wang, "Deep vision multimodal learning: Methodology, benchmark, and trend," *Appl. Sci.*, vol. 12, no. 13, p. 6588, Jun. 2022.

[10] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 1–20, Oct. 2023.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 8748–8763.

[12] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[13] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "PointCLIP: Point cloud understanding by CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8542–8552.

[14] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 976–980.

[15] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. J. Autom. Comput.*, vol. 18, no. 3, pp. 351–376, Jun. 2021.

[16] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 17612–17625.

[17] Q. Jiang, C. Chen, H. Zhao, L. Chen, Q. Ping, S. D. Tran, Y. Xu, B. Zeng, and T. Chilimbi, "Understanding and constructing latent modality structures in multi-modal representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7661–7671.

[18] L. Ruan, A. Hu, Y. Song, L. Zhang, S. Zheng, and Q. Jin, "Accommodating audio modality in CLIP for multimodal processing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 9641–9649.

[19] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning robust audio representations from clip," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4563–4567.

[20] X. Zhan, H. Liu, Q. Li, and A. B. Chan, "A comparative survey: Benchmarking for pool-based active learning," in *Proc. Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 4679–4686.

[21] S. Liu, K. Tang, and X. Yao, "Generative adversarial construction of parallel portfolios," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 784–795, Feb. 2022.

[22] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 721–725.

[23] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[24] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5066–5074.
[25] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2018, pp. 135–150.
[26] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "MultiMAE: Multi-modal multi-task masked autoencoders," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 348–367.
[27] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 1–12.
[28] K. Al-Hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: A tutorial and survey," *Vis. Comput. Ind., Biomed., Art*, vol. 6, no. 1, p. 14, Jul. 2023.
[29] M. Tschannen, B. Mustafa, and N. Houlsby, "CLIPPO: Image-and-Language understanding from pixels only," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11006–11017.
[30] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 6074–6082.
[31] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, May 2021, pp. 2302–2310.
[32] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 2776–2784.
[33] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15878–15887.
[34] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18156–18165.
[35] H. Maheshwari, Y.-C. Liu, and Z. Kira, "Missing modality robustness in semi-supervised multi-modal semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 1009–1019.
[36] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2023, pp. 19730–19742.
[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[38] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," 2021, *arXiv:2104.01778*.
[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[40] E. Brynjolfsson, "The Turing trap: The promise & peril of human-like artificial intelligence," *Daedalus*, vol. 151, no. 2, pp. 272–287, May 2022.
[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[42] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3498–3505.
[43] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.
[44] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 1, 2009. '
[45] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*.
[46] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.
[47] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia (ACM-MM)*, Nov. 2014, pp. 1041–1044.
[48] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (BiT): General visual representation learning," in *Proc. 16th ECCV*, Jan. 2020, pp. 491–507.
[49] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[50] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
[51] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "ESResNe(X)t-fbsp: Learning robust time-frequency transformation of audio," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.
[52] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.
[53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[55] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, Jun. 2005, pp. 65–72.
[56] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.
[57] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Proc. SLTU*, Jan. 2008, pp. 63–68.



SANGYEON CHO (Graduate Student Member, IEEE) received the B.S. degree in software engineering from Kookmin University, Seoul, Republic of Korea, in 2023. He is currently pursuing the M.S. degree with Chung-Ang University. His research interests include natural language processing, representation learning, and various multi-modal learning.



JANGYEONG JEON (Graduate Student Member, IEEE) received the B.S. degree from the Department of Big Data Statistics, Cheongju University, Cheongju, Republic of Korea, in 2022. He is currently pursuing the M.S. degree in artificial intelligence with Chung-Ang University, Seoul, South Korea. His main research interest includes natural and various multi-modal large language models.



MINGI KIM received the B.S. degree in business administration from Chung-Ang University, Seoul, Republic of Korea, in 2023, where he is currently pursuing the M.S. degree. His research interests include fairness, natural-language processing, and various multi-modal learning.



JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, Republic of Korea, in 2015, 2017, and 2021, respectively. He was a Postdoctoral Research Associate with the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He is currently an Assistant Professor with the Department of AI, Chung-Ang University, Republic of Korea. His research interests include visual-language reasoning, visual question answering, and various video-based problems.