The Journal of Supercomputing (2025) 81:1114
https://doi.org/10.1007/s11227-025-07473-7

RESEARCH

Saffe: Multimodal Model Composition with Semantic-Alignment Fusion of Frozen Encoders

Maithri Kulasekara[1,3] · Juan F. Inglés-Romero[4] · Baldomero Imbernón[2] · José L. Abellán[1]

Accepted: 16 May 2025
© The Author(s) 2025

Abstract
Transformer-based multimodal models often require expensive, full-model training on task-specific all-modality datasets to achieve high accuracy on targeted downstream tasks. To reduce this significant cost, we introduce SAFFE, a methodology for building accurate, task-specific multimodal models with minimal training, using only standard GPU hardware. SAFFE leverages off-the-shelf, pre-trained, frozen unimodal encoders for each input modality (e.g., text, image, or audio) and connects them through a lightweight, trainable component called the FusionAlign Module (FAM). FAM is a bottleneck mid-fusion neural network, trained on the target dataset to align the outputs of the independently pre-trained unimodal encoders. This approach eliminates the need for end-to-end training while maintaining strong accuracy for the downstream task. As a proof of concept, we validate SAFFE on image retrieval and language understanding tasks. SAFFE-derived models outperform state-of-the-art multimodal systems on datasets such as CIFAR-10, ImageNet-100, and COCO, achieving competitive results with significantly fewer trainable parameters and training time.

Keywords Multimodal fusion · Frozen model · Transformer encoder · Decoder · Mid-fusion

1 Introduction and motivation

In recent years, multimodal learning has gained significant attention for its ability to process various input modalities concurrently, such as text, images, and audio, to enhance the prediction accuracy of downstream tasks, as integrating multiple modalities provides more comprehensive semantic information compared to approaches relying on single modalities in isolation [1–14].

Transformer-based multimodal models attain enhanced prediction accuracy, exceeding the performance of traditional convolutional and recurrent neural

Extended author information available on the last page of the article

Published online: 07 July 2025

network-based multimodal techniques [1, 8, 9, 15–21]. The core architecture of Transformer-based multimodal models generally comprises both encoder and decoder modules [22]. Furthermore, this new type of multimodal model is highly versatile. For instance, by combining text and visual inputs, they can support a wide range of tasks, such as image-text retrieval [18, 20], scene analysis [18], image segmentation [23, 24], diffusion models [25, 26], scene graph generation [27, 28], visual question answering (VQA) [21], and generating responses that seamlessly integrate textual and visual information.

Nonetheless, developing an effective multimodal model that excels at modern downstream tasks presents several challenges. One major hurdle is the computationally intensive end-to-end training process, which often relies on cloud infrastructure using up to several thousand compute nodes and can incur costs exceeding millions of dollars [22, 29]. In addition, these models require large-scale datasets—often on the order of petabytes—across all modalities [1, 3, 8, 9, 30], and may contain up to trillions of parameters (e.g., ChatGPT, Google Gemini, or GPT-J [31]). The training process must also address discrepancies in learning dynamics across modalities, account for diverse noise patterns (where certain modality streams contribute more task-relevant information than others), and incorporate specialized input representations [32, 33].

Representative examples of state-of-the-art (SOTA) multimodal models include CLIP [9], OpenCLIP [34], CoCa [8], Laion [30], and Chameleon [29], which focus on image-text fusion. Additionally, models such as VATT [1] and VALOR [35] integrate text, image, and audio modalities. However, training these large-scale multimodal models to achieve high accuracy is extremely costly, as it requires end-to-end training across all model parameters—which can comprise over a billion parameters. As a result, adapting these models to application-specific downstream tasks typically demands data-center scale compute resources. For instance, Chameleon was trained using 3072 GPUs over a total of 428,207 GPU hours, CLIP was trained for 2 weeks on 256 GPUs, and VATT utilized 256 TPU v3 chips over a period of 3 days.

Efforts to reduce training expenses have led to the development of modality fusion frameworks, such as Flamingo [18], UniT [36], BLIP [20], MAGMA [21], and FROMAGe [19], which build multimodal models by combining existing pre-trained unimodal encoders and decoders. While these frameworks represent progress in reducing training costs, they introduce new components that are closely tied to their specific encoders and/or decoders. However, these customized components, such as the interleaved connection between encoders and decoders, build new attention mechanisms and introduce new tokens [18, 20, 21, 36, 37] preventing the utilization of off-the-shelf pre-trained models. This dependency on customized integration still requires significant computational resources. For example, BLIP requires 8 days of training on eight A100 GPUs, UniT necessitates 3 days on eight V100 GPUs, Flamingo needs 15 days on TPUv4, and MAGMA demands 1.75 days on sixteen A100 GPUs, leading to substantial training expenditures (see Section 4.2 for further details).

To address the challenge of building high-accurate multimodal models on evolving downstream tasks and specific datasets, while minimizing the high training

costs, we introduce the Semantic-Alignment Fusion of Frozen Encoders (SAFFE) methodology.

SAFFE leverages off-the-shelf, pre-trained, and frozen unimodal encoders for each input modality, such as text, image, or audio. These encoders, trained on large, general-purpose datasets and widely available from leading AI companies such as OpenAI, Meta, Google, LangChain, and Hugging Face, are kept fixed during training. This approach reduces computational costs and eliminates the need for end-to-end model training. The use of pre-trained unimodal encoders in our methodology to compose task-specific multimodal models offers two key advantages. First, it eliminates the need for costly end-to-end training across all modalities (e.g., text, image, and audio), as seen in SOTA models like VATT. Second, SAFFE benefits from zero-shot learning (ZSL) capability, enabling frozen encoders to work with other modalities without prior training, and enhances prediction accuracy across different datasets.

However, no existing methodology effectively integrates independently pre-trained unimodal encoders in a modality-agnostic manner to build a unified multimodal understanding framework for downstream tasks. To address this gap, we introduce the FusionAlign Module (FAM)—a lightweight, trainable bottleneck mid-fusion neural network. FAM is trained on the target dataset to align and combine the output representations of frozen unimodal encoders, extracting modality-specific features while enabling seamless integration across modalities. This design eliminates the need for computationally expensive, end-to-end joint training over all modality-paired datasets. SAFFE leverages FAM to operate in a truly modality-agnostic fashion, supporting a wide range of downstream tasks by flexibly incorporating diverse pre-trained encoders. It maintains a unified architecture for all modalities, simplifying both the training and deployment processes for efficient multimodal fusion.

At a high level, our SAFFE methodology involves two sequential stages for composing multimodal models tailored to specific downstream tasks and datasets (see Figure 1). First, we capitalize on the wealth of existing downloadable off-the-shelf frozen pre-trained components such as vision encoders, linguistic encoders, or audio encoders. Second, we demonstrate that a lightweight training strategy, focused solely on aligning the semantics of the output vector spaces of each individual pre-trained component through the target all-modality datasets, can achieve high prediction accuracy for the corresponding downstream task. In this approach, training is confined to the newly introduced FAM unit. It consists of lightweight, fusion-specific components that leverage bottleneck mid-fusion while leaving the original pre-trained frozen components untouched (see Figure 2). As a result, SAFFE effectively lowers both computational costs and training time for the customized multimodal model while delivering high prediction accuracy for the specified downstream task.

To validate the effectiveness of the SAFFE methodology, we conduct an extensive set of experiments in a bimodal setting involving image and text modalities. These experiments evaluate various configurations for composing multimodal models tailored to downstream tasks such as image retrieval and language understanding. For benchmarking, we use widely recognized datasets, including ImageNet100 [38], CIFAR-100 [39], and COCO [8], and compare SAFFE against SOTA multimodal

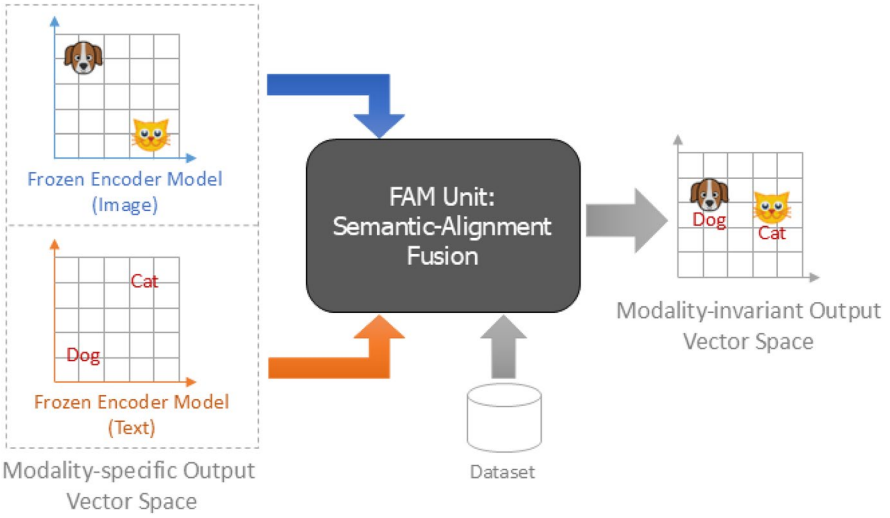


Fig. 1 Overview of our SAFFE methodology. To build a customized multimodal model (a SAFFE-derived model) for a given downstream task (e.g., image retrieval), different per-modality frozen components depending on the target input modalities (e.g., text and images) are taken from an official repository (e.g., Hugging Face and Kaggle). Then, these frozen components are connected with a FAM unit (details in Figure 2). FAM is trained with the task-specific dataset for inexpensive semantic alignment of every modality-dependent frozen encoder, producing a modality-invariant output vector space for accurate model deployment on the downstream task

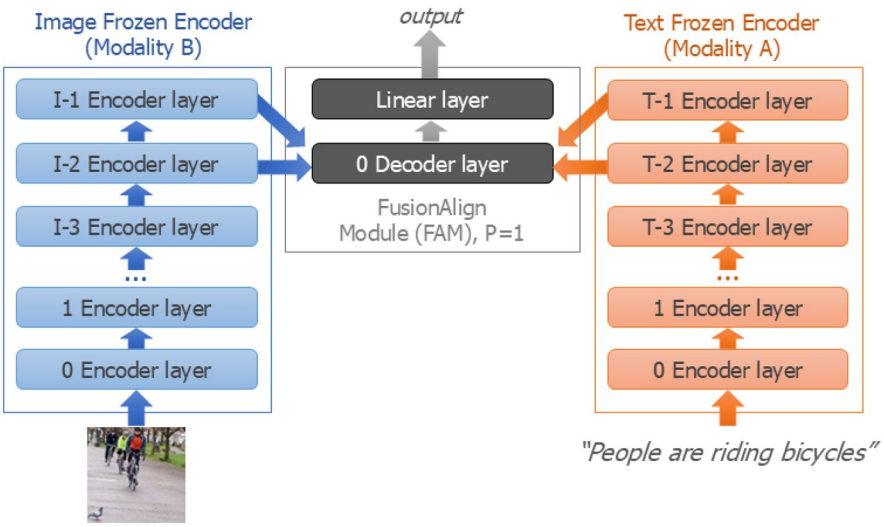


Fig. 2 A SAFFE-derived bimodal model with frozen image and text encoders, along with a high-level architecture overview of a FAM unit (details in Section 3.2.1). The vertical arrows indicate the outputs from the encoders, while the other arrows are the weights (K, Q, V)

models and SOTA fusion techniques. To support further research and development, we release SAFFE as an open-source tool [1] for the research community.

The key contributions of this work include:

- We propose SAFFE, a methodology for flexible composition and efficient training of multimodal models to achieve high prediction accuracy on targeted downstream tasks. SAFFE enables users to select and integrate off-the-shelf, frozen encoders for each input modality, tailored to the specific requirements of the target task. The fusion of these encoders is facilitated by FAM, a lightweight, trainable bottleneck mid-fusion neural network, which is trained exclusively on the task-specific dataset. This design significantly reduces the full-model training costs typically associated with SOTA multimodal systems, making it practical to build accurate models using standard, single-node GPU-based hardware. In particular, all experiments in this work were conducted on a single Nvidia RTX 3060 GPU, with the longest run completing in approximately 2.5 days.
- SAFFE-derived bimodal models for image retrieval tasks achieve optimal mean Average Precision (mAP) by training FAM using only the final two layers of each pre-trained encoder. This contrasts with SOTA approaches, which typically apply full pairwise attention across all Transformer-based encoder layers using the entire task-specific dataset. FAM achieves near-optimal performance when trained on just 50% of the dataset while requiring fewer than 30 epochs to reach maximum mAP. This leads to significant reductions in both computational cost and training time.
- By effectively composing SAFFE-derived models, we outperform bimodal architectures in both mAP and computational efficiency. Our model surpasses Flamingo [18] by 1.3% mAP on the COCO dataset while requiring three orders of magnitude fewer trainable parameters (70B vs. 67 M). Furthermore, by training only the FusionAlign Module (FAM), SAFFE achieves higher zero-shot accuracy than frozen SOTA multimodal models such as ZLaP [40] on CIFAR-10 and CIFAR-100, with less than half the trainable parameters (151 M vs. 67 M). SAFFE also enables rapid integration of novel linguistic concepts through FAM alone. Compared to the SOTA Meta-Learning framework [41], our approach consistently yields significantly higher prediction accuracy after just a single training epoch, achieving improvements ranging from 21% to 56%.

2 Background and related work

In this section, we discuss the fundamental concepts underpinning SAFFE, examine the prevailing techniques for feature fusion in existing models, and highlight the specific limitations relevant to our application example that focuses on text and image modality fusion.

[1] *https://github.com/CAPS-UMU/SAFFE.git.*

2.1 Transformers

The Transformer architecture has become a cornerstone in natural language processing (NLP) due to its ability to handle various tasks with high accuracy and efficiency. The linguistic transformer models, BERT [42], SBERT [43], and RoBERTa [44], are frequently highlighted for their robust performance across multiple NLP tasks. The Transformer architecture has found significant applications in other areas of artificial intelligence, notably in computer vision [45], speech recognition [46], Anomaly Detection [47], video classification [48] and Video-Audio-Text Transformer (VATT) [1], to name a few.

2.1.1 Components of transformers

The primary constituents of a Transformer architecture are the encoders and decoders. These components differ markedly, with each fulfilling unique functions in the manipulation and generation of sequential data [22]. Encoders serve as gateways to contextual information, converting it into a sequence of vector representations that capture the input's semantic and syntactic attributes. The decoder produces the result by combining the output from the encoder with additional vector representation information from another modality [49]. The main building blocks of these popular Transformer-based models consist of a sequence of Multi-Headed Self-Attention (MSA), MCA, and feedforward neural networks. In encoder–decoder architectures, the last encoder layer output generated by the encoder is conventionally utilized by the decoder via a MCA mechanism, thereby enabling the decoder to concentrate on pertinent segments of the input sequence. This interaction between the encoder and decoder is crucial for generating contextually relevant outputs, allowing the model to produce more coherent and accurate responses based on the input it receives.

2.1.2 Types of transformer-based models

Language Models: The Transformer text encoder with BERT (Bidirectional Encoder Representations from Transformers) [42] is a groundbreaking model in NLP, that leverages the Transformer architecture to understand the context of words in a text. BERT's innovation lies in its ability to process text bidirectionally, meaning it considers the context from both the left and right of a word, which enhances its understanding of linguistic nuances. This capability has made BERT a preferred choice for various NLP tasks, including text classification, sentiment analysis, and information retrieval. Furthermore, BERT's pre-training on vast amounts of text data allows it to capture a wide range of linguistic patterns, making it highly effective in generating contextual embedding vectors that can be fine-tuned for specific applications [42].

Vision models: The Vision Transformer (ViT) [45] encoder is a transformative architecture in image processing, leveraging the MSA mechanism to handle image data effectively. It divides an image into patches, processes these patches as tokens, and uses a Transformer encoder to extract features. This approach allows ViT to

capture long-range dependencies and contextual information across the image. The architecture's functionality is enhanced by positional encoding, which helps maintain the spatial relationships between patches. The ViT model employs MSA to process the sequence of patches. This mechanism enables the model to focus on different parts of the image simultaneously, capturing complex patterns and relationships [9, 10, 30]. At the end of the process, the ViT encoder produces a significant feature vector for the patches, which is essential for applications such as image classification, image segmentation, and scene understanding [8, 9, 18, 20].

2.2 Taxonomy of multimodal fusion strategies

Multimodal fusion techniques are essential for integrating diverse data types (e.g., image, text, and audio). This creates more robust representations, improving the performance of Transformer-based models across real-world applications by enabling contextually richer outputs [6].

As illustrated in Figure 3, these fusion techniques are divided into early fusion, mid-fusion, and late fusion, each possessing unique approaches and consequences. Early fusion merges multiple data modalities right at the input level, enabling the model to benefit from integrated data from the outset [24, 29, 50]. The current state of the art indicates that early fusion approaches exhibit reduced precision due to several factors, including discrepancies in learning dynamics between modalities, variations in noise topologies, and distinct input representations among modalities [32, 33]. Late fusion aggregates the outputs of distinct models that have been trained on separate modalities, frequently leading to a more generalized outcome [5, 51]. Bidirectional learning and shared feature representation are not possible with the late fusion method [5, 52]. However, mid-fusion techniques have shown promise in addressing these challenges by leveraging the strengths of each modality while minimizing their weaknesses, thereby enhancing overall performance in multimodal tasks [53]. These methods leverage the strengths of Transformers, such as their ability to model long-range dependencies and process multiple modalities within a unified framework [36]. The primary advantages linked to mid-fusion techniques that

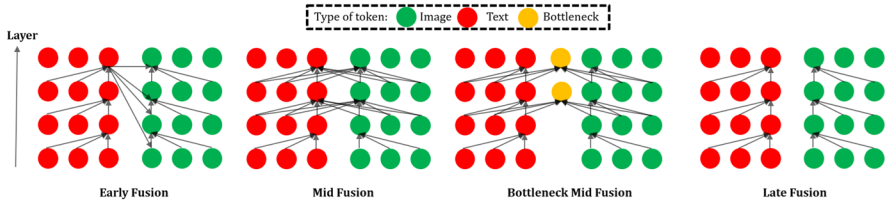


Fig. 3 **Multimodal Fusion Strategies**. Early Fusion image (left): Cross-modal information is exchanged at the initial input stage of the model [24]. Mid-Fusion image (middle, left): Cross-modal interactions are implemented using pairwise MCA mechanisms across subsequent layers. Bottleneck Mid-Fusion in SAFFE image (middle, right): We leverage "Bottleneck Mid-Fusion," which restricts the flow of attention within a given layer using the proposed FAM units. Late Fusion image (right): Cross-modal information is exchanged only after the classification stage, with no interaction between modalities during earlier stages. Figure adapted from [3]

concatenate two modalities of data include the provision of a unified architecture that can handle both types of data, allowing for the seamless integration of multimodal information. This is particularly beneficial in tasks such as visual question answering and sentiment analysis, where understanding the context from both modalities is crucial [53]. Moreover, bidirectional learning is advantageous for tasks requiring an understanding of the relationships among modalities, such as in text-image fusion processed by image-to-text and text-to-image generation [3, 18, 20]. This approach reduces the need for separate models for each task, simplifying design and improving efficiency.

MBT [3] proposes a novel architecture for audiovisual fusion that restricts the flow of cross-modal information through "Tight Bottleneck Fusion" in the model (see Figure 3). This mid-fusion model focuses on condensing the most relevant information from each modality and sharing only what is necessary with the other modality. By doing so, the model avoids the quadratic scaling costs of full pairwise attention and achieves performance gains with fewer computational resources. However, the model proposed by MBT consists of 12 encoders coupled with 4 decoders. This comprehensive end-to-end training approach incurs significant computational expenses.

2.3 Advanced image-text fusion techniques

Recent advancements in vision-language models (VLMs) have led to the development of innovative frameworks such as BLIP [20], Flamingo [18], and UniT [36], each employing unique architectures for effective image-text fusion. BLIP utilizes a multimodal mixture of Encoder–Decoder (MED) architecture, enhancing tasks such as image-text retrieval and caption generation through pre-trained weights from ViT and BERT, while also sharing parameters between encoders and decoders to optimize performance. Flamingo, on the other hand, introduces novel techniques for few-shot learning by bridging vision-only and language-only models, employing a Perceiver Resampler to condition a frozen language model with visual tokens derived from interleaved visual and textual data. Similarly, UniT adopts a transformer architecture with a mid-fusion approach, facilitating concurrent learning across various tasks, and integrates multiple encoders and decoders for joint training, albeit at high computational costs.

FROMAGe [19] grounds a frozen autoregressive LLM in the visual domain, utilizing a pre-trained visual model to extract embeddings from images and integrate them into the LLM's input space. This model introduces a special [RET] token to improve image retrieval based on text input, though it exhibits a bias toward generating regular text tokens due to its text-centric pre-training. In contrast, the Frozen framework [52] extends the soft-prompting technique of prefix tuning to enable multimodal few-shot learning, focusing on open-ended image interpretation while primarily training on captioning tasks. This framework employs a pre-trained linguistic model and a visual encoder derived from NF-ResNet-50. Pang et al. [54] introduce an innovative approach by incorporating frozen Transformer blocks from pre-trained LLMs into visual encoding layers, proposing the "information filtering hypothesis" that these blocks enhance the

identification of significant visual tokens. This model employs the LLaMA [55] modality alongside VIT [45], focusing on image-to-text tasks. The SIGLIP [37] model features a novel pairwise sigmoid loss function designed for image-text pairs, enhancing the training efficiency of multimodal models. It was trained using a Base CLIP architecture with a batch size of 4000 and a Large LiT model with a batch size of 20,000, involving 1.8 billion parameters over 5 days on 32 TPU-v4 chips. This innovative approach aims to improve the alignment between visual and textual data, advancing language-image pre-training capabilities. Lastly, the MAGMA [21] technique trains a linear layer to project image representations into the language model's spatial domain, facilitating vision-language task assessments without altering other model parameters. This late fusion mechanism employs a pre-trained image encoder alongside text decoders and undergoes training for 15,000 iterations.

Collectively, these models exemplify the efficacy of Transformer-based architectures in enhancing multimodal tasks; however, the encoder architectures of these models are specifically tailored with distinct components. Due to these new components, off-the-shelf encoder models are inapplicable in their original form, thereby requiring high computational resources to effectively train them.

3 SAFFE methodology

3.1 General procedure

As introduced in Section 1, our SAFFE methodology for multimodal model composition targeting a specific downstream task consists of two main phases (see Figure 1). In the first phase, the chosen frozen encoders are suitable for particular downstream tasks pertinent to the corresponding input modalities. For example, in a bimodal scenario combining image and textual data, this involves choosing extensively trained linguistic and vision-based off-the-shelf encoders with zero-shot classification capabilities, such as those from the Hugging Face, Kaggle repository. In the next phase, after connecting the selected input encoders to our FAM unit, we train the FAM to achieve semantic alignment and fine-tune the composed multimodal model for the downstream task, as shown in Figure 2. As we will explain below, the complexity of training FAM depends on the relationship between the encoders. This relationship forms the basis of our classification into partially-aligned and non-aligned encoders.

3.2 A SAFFE-derived bimodal model architecture

For the sake of generality in our bimodal scenario, we termed both modalities as *ModalityA* and *ModalityB*. The frozen models are associated with *ModalityA* and *ModalityB* and are designed to accept inputs relevant to their corresponding modalities, with the encoders for *ModalityA* and *ModalityB* generating hidden states in conjunction with pre-trained weights specifically intended for the design of FAM.

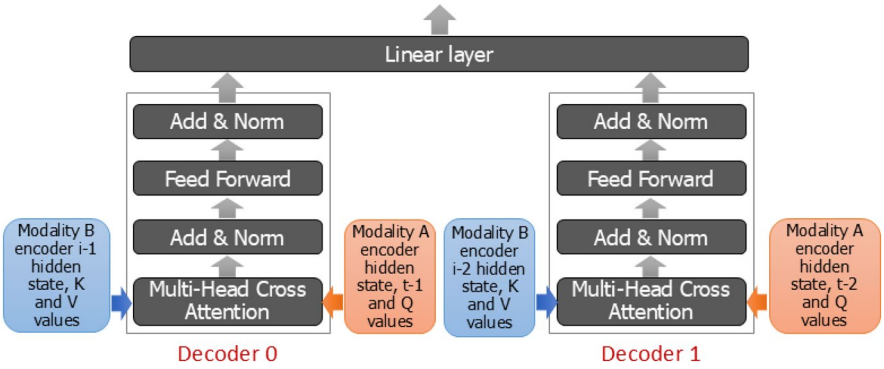


Fig. 4 Architecture of the most complex FAM unit used in our experiments (Section 4). FAM functions as a mediating element that connects the *ModalityB* and *ModalityA* feature vectors. The Keys (*K*) and Values (*V*) within these layers are sourced from the *ModalityB* features, while the Queries (*Q*) are generated from the *ModalityA* inputs. The final two successive layers of *K*, *Q*, and *V* undergo concatenation with two decoders, followed by a linear layer that densifies them, resulting in the generation of an output feature vector

3.2.1 The fusion align module (FAM) unit

As illustrated in Figure 2, our methodology involves training a decoder layer utilizing the hidden state (h^{i}) derived from a *ModalityB* encoder (e.g., an image frozen encoder) that comprises a total of I layers ($(E^{i}) : i = 0...., I-1$), which will be integrated as inputs into a fusion with the hidden state (h^{t}) of a frozen *ModalityA* encoder (e.g., a text frozen encoder) that consists of a total of T layers ($(E_{t}$): $t = 0..., T-1$), aimed at creating a model for a *ModalityA* and a *ModalityB* content fusion application. The (h^{i}) and (h^{t}) states, along with their corresponding K, Q, and V values, are transmitted to the decoder module; an explanation of the decoder module is provided in Subsection 3.2.2. The process of generating the decoder output involves a linear layer that produces the ultimate pooled output.

Transformers require full pairwise attention to enable the encoder–decoder relationship [22]; however, the MBT manuscript makes the observation that extensive pairwise attention across all layers is unnecessary; hence, they propose the bottleneck mid-fusion units [3]. Building on these insights, we designed the structure of our FAM unit—Figure 4 shows the most complex 2-decoder FAM unit used in our experiments—to achieve semantic alignment-driven fusion between *ModalityA* and *ModalityB* using frozen encoders. FAM employs the mid-fusion technique, utilizing off-the-shelf pre-trained encoders to bypass the need for comprehensive training of both encoders and decoders.

As we can see, *ModalityA* and *ModalityB* inputs are initially processed with the frozen unimodal encoder to extract the specific h^{i} and h^{t} states, which are subsequently passed through a decoder model to enhance the interaction among modalities. The core concept is to merge a small set of the later few ($t = T-1, T-2$ and $i = I-1, I-2$; T and I are the numbers of encoders for *ModalityA* and *ModalityB*, respectively) hidden states of the frozen encoders using MCA. Due to this

mechanism, our methodology diverges from other prevailing techniques that predominantly depend on comprehensive attention throughout all layers [3, 20], which necessitates substantial computational resources, thus facilitating a more efficient and focused fusion process [54]. The hidden state of the *ModalityA* and *ModalityB* encoder encapsulates the information pertaining to *ModalityA* and *ModalityB*, respectively. The suggested decoder model integrates this information and produces a cohesive output that seamlessly merges both modalities, thereby enhancing the model's overall effectiveness for downstream tasks.

Partially-aligned vs. Non-aligned encoders: As we will demonstrate in Section 4, our FAM unit architecture has demonstrated effectiveness in a bimodal scenario, utilizing image and text frozen encoders trained on different datasets and originating from different model owners. For instance, an image frozen encoder such as openai/clip-vit-base-patch32 and a text frozen encoder such as sentence-transformers/all-mpnet-base-v2 is designated as $\text{SAFFE}_{\text{Non-aligned}}$.

Nonetheless, when the semantic alignment of both modalities is less challenging, i.e., the frozen encoders belong to the same frozen multimodal model or they have already been trained with similar datasets (for instance, a frozen image encoder such as openai/clip-vit-base-patch32, and a text frozen encoder such as openai/clip-vit-large-patch14), it is designated as $\text{SAFFE}_{\text{Partially-aligned}}$. Our experiments reveal that it is not necessary to train the decoder part of our FAM component but only its linear layer to achieve high prediction accuracy.

3.2.2 The architecture of the FAM's decoder module

ModalityB frozen models produce a diverse array of latent components, including hidden states, Queries (Q), Keys (K), and Values (V), each characterized by its specific weights, whereas *ModalityA* frozen models generate T distinct hidden states, Q, K, and V, similarly endowed with their weights. We have devised a training methodology employing pre-trained frozen transformers' constituents. We represent the output of a *ModalityA*, T encoder layer as $E^{\text{T}+1} = Transformer(E^{\text{T}})$, and the output of a N decoder layer as $D^{\text{N}+1} = Transformer(D^{\text{N}})$.

The SAFFE methodology is not commutative; thus, we regard *ModalityA* as an output modality type for clarification. More specifically, *ModalityA* generates the Q, while *ModalityB* supplies the K and V utilized within the MCA methodology, as shown in Equation 1.

$$\begin{aligned} &MCA(ModalityA, ModalityB) \\ &= Attention(W^{\text{Q}}ModalityA, W^{\text{K}}ModalityB, W^{\text{V}}ModalityB) \end{aligned} \tag{1}$$

Where the decoder utilizes the MCA, which assimilates the output generated by the encoder E^{T}. D^0 MSA is derived from the *ModalityA* encoder as shown in Equation 2. The decoder unit employs MSA, utilizing the K, Q, and V values along with the weight parameters from both encoders to compute the MCA as shown in Equations 3 and 4. Layer normalization is denoted as LN.

$$u' = MSA(LN(h^{t-1})) + h^{t-1} \tag{2}$$

$$Z' = MCA(ModalityA, ModalityB) + u' \tag{3}$$

$$D^{N} = LN(Z') \tag{4}$$

Ultimately, the outputs generated by these decoders are subsequently fed into the linear concluding layer, where they undergo a succession of transformations to yield the intended output sequence.

Reusing the weights of the frozen encoder for the FAM's decoder: The fusion of non-aligned encoders is facilitated by the decoder module. This module comprises both a MCA mechanism and a MSA mechanism. Instead of establishing new weights for the FAM's decoder, we derive them from the respective weights of the original encoder's subsequent layer (see Figure 4). This weight-sharing approach improves the model's effectiveness, facilitating superior generalization across various tasks while preserving a unified representation of both modalities.

3.3 Applicability and constraints

As in any methodology, SAFFE has certain considerations and constraints for its effective applicability on multimodal downstream tasks.

1. **FAM unit:** The design of this component is critical for effectively fusing modality-dependent frozen encoders for achieving the highest model performance in a target downstream task. Users must explore the design space of the architectural components in FAM, especially determining the number and type of decoder layers. To assist with this exploration in the context of bimodal image retrieval and language understanding tasks, users can refer to the insights summarized in our experimental evaluation (Section 4). For other downstream tasks involving different modalities and/or other task-specific datasets, the FAM unit will require redesign. Our future work will focus on expanding this development roadmap to further demonstrate the effectiveness of SAFFE in these other contexts.
2. **Frozen encoders**: The selection of unimodal encoders plays a crucial role in the performance of SAFFE-derived models. A frozen encoder is required for each modality, and each must provide access to the Transformer encoder's output, as well as its corresponding KQV values and weights, as needed by the FAM unit. Additionally, all encoders must produce outputs with matching vector dimensions to enable efficient execution of multi-head self-attention (MSA), cross-attention (MCA), and seamless integration. Our evaluation in Section 4 demonstrates the impact of different frozen encoders in bimodal scenarios, offering practical guidance for model composition. For broader deployments involving other modalities, recent research on vector dimension alignment [56–58] can support further optimization.

3. **Multimodal Datasets:** Training the FAM unit within a SAFFE model requires a multimodal dataset, and both the quality and quantity of this data are critical to the method's effectiveness. As shown in our evaluation (Section 4.2.2), strong semantic alignment can be achieved using only 50% of the dataset, with performance close to the optimal. This demonstrates the potential of SAFFE in scenarios where multimodal data is scarce for new downstream tasks.

4 Evaluation and results

In this section, we demonstrate the benefits of our SAFFE methodology applied to a bimodal case study composed of text and image modalities with image retrieval and language understanding tasks. To this end, we initially describe the datasets employed for the training and evaluation of multimodal fusion (Section 4.1.1). Then, we delve into the experimental setup used to carry out our experiments (Section 4.1.2). Finally, we analyze our experimental results (Section 4.2).

4.1 Implementation details

4.1.1 Datasets

In our experiments, we first use three distinct datasets with increasing complexity: Dogs vs. Cats [59], CIFAR-10 [39], and ImageNet100 [38]. Additionally, we use CIFAR-100 [39] and COCO [8], for a comparative analysis of our SAFFE-derived bimodal models against SOTA approaches. All relevant dataset specifications are provided in Table 1.

4.1.2 Experimental setup

Our SAFFE methodology and derived bimodal models are implemented in PyTorch [64]. The off-the-shelf vision frozen encoder models are initialized using the ViT architecture and are pre-trained [9, 30, 65]. The off-the-shelf frozen encoder models dedicated to textual data are initialized using the BERT architecture and are also pre-trained [9, 30, 43, 65]. All the off-the-shelf frozen encoder models used in our evaluation are acquired from Hugging Face's Transformers [66] library.

Table 1 Datasets used to evaluate the bimodal case study of our SAFFE methodology

Datasets	#Classes	Training Set	Testing Set	Image Size
Dogs vs. Cats [59]	2	22,500	2500	100×100 to 2000×1000
CIFAR-10 [60]	10	50,000	10,000	32×32
ImageNet100 [61]	100	100,000	5000	469×387
CIFAR-100 [62]	100/20 superclasses	50,000	10,000	32×32
COCO [63]	80	117,200	5000	640×480

For our semantic-alignment fusion experiments, we consider the following configurations (details in Table 2). In case of the $\text{SAFFE}_{\text{Partially-aligned}}$, we examine the image encoder specified as openai/clip-vit-base-patch32 alongside the text encoder designated as openai/clip-vit-large-patch14 [9]. These CLIP frozen models have been trained by the OpenAI team using a dataset comprised of 400 million image-text pairs sourced from the Internet [9]. In this case, as detailed in Section 3.2.1, a simple FAM structure consisting of only a linear layer (refer to Figure 4) is sufficient to achieve high prediction accuracy (the experimental results are shown below). On the other hand, in the case of the $\text{SAFFE}_{\text{Non-aligned}}$, we use the same image encoder as before, openai/clip-vit-base-patch32 [9], but the text encoder is unrelated to the image encoder, selecting sentence-transformers/all-mpnet-base-v2 [43]. It utilized the pre-trained microsoft/mpnet-base model and was trained on a dataset comprising 1 billion sentence pairs. This more challenging semantic alignment of both types of encoders necessitates a more complex FAM unit. In this scenario, our FAM unit employs the full architecture depicted in Figure 4. It includes a trainable decoder layer for each modality-specific stack of encoder layers, which is responsible for generating outputs for the trainable linear layer.

For the training procedure of FAM in our SAFFE-derived bimodal models, we employ the weighted Adam [67] optimization algorithm, utilizing a standard learning rate of $2e^{-4}$ [20]. The Adam optimizer effectuates updates to the model parameters based on the gradients derived from the task-specific losses. During the inference phase, we employ a system output vector as the modality representation and compute cosine similarity [9] to evaluate the mAP value for the image retrieval task. Additionally, we employ a batch size of 30 and set the number of epochs to 50 for all experiments. These hyperparameters achieve the highest prediction accuracy for the produced SAFFE-derived multimodal models under evaluation. Our training and prediction experiments require only commodity hardware; specifically, we use an NVIDIA GeForce RTX 3060 GPU with 12.74 TFLOPS of computational power and 12 GB of GDDR6 memory.

4.2 Experimental results

In this section, we examine the impact of various architectural choices in SAFFE-derived multimodal models. The analysis focuses on the extent to which the contributions of encoder layers influence the functionality of FAM units in achieving optimal output, as well as the effect of the volume of training data on accuracy.

Table 2 SAFFE-derived models with their associated frozen encoders and features. *EL*=Number of encoder layers and *ED*=Embedding Dimension

Model	Text Encoder	Image Encoder	EL	ED
$\text{SAFFE}_{\text{Partially-aligned}}$	openai/clip-vit-large-patch14	openai/clip-vit-base-patch32	12	768
$\text{SAFFE}_{\text{Non-aligned}}$	sentence-transformers/all-roberta-large-v1	openai/clip-vit-base-patch32	12	768

Ultimately, this analysis pertains to the accuracy associated with the conceptual integration of visual and linguistic elements within the framework of FAM.

4.2.1 Analysis with fusion strategies

In general, the FAM unit can take as input P encoder hidden states from a single modality, where P can range from one to the total number of encoder layers in the frozen model. In our $\text{SAFFE}_{\text{Partially-aligned}}$ and $\text{SAFFE}_{\text{Non-aligned}}$ models, we conduct experiments with $P = 1$, a single encoder, and $P = 2$, two consecutive encoders as shown in Figure 5. Note that in the latter case ($P = 2$), only the last two consecutive layers from both encoder layers are used to train the FAM unit.

For the experimental design, among all the 12 available encoder layers per modality (see Table 2), we conduct experiments by selecting a subset of them. The purpose of these experiments is to evaluate the mAP values of the FAM unit in the bimodal model using the chosen encoder layers for each modality, derived from the frozen pre-trained models. Specifically, we select five experimental cases for both scenarios: 0, 2, 5, 8, and 11 encoder layers for the $\text{SAFFE}_{\text{Partially-aligned}}$ model, and 0, 2, 5, 8, and 10 encoder layers for the $\text{SAFFE}_{\text{Non-aligned}}$ model for $P = 1$ case. Note that each of these values represents the encoder layer identifier used to train the FAM unit. For instance, in the experiment with "5," we use the fifth encoder layer. Additionally, in the non-aligned case, as illustrated in Figure 2, since we use the K, Q, and V weights from the next encoder layer, the maximum encoder layer identifier we can test is "10." These encoders are represented according to their respective encoder numbers on the y-axis ("Fusion Layer") of Figures 6a through 8c.

In the case of $P = 2$, the $\text{SAFFE}_{\text{Partially-aligned}}$ model uses 0, 2, 4, 6, and 10 encoder layers with consecutive layers, represented graphically as $0 + 1, 2 + 3, 4 + 5, 6 + 7$, and $10 + 11$. Similarly, in the $\text{SAFFE}_{\text{Non-aligned}}$ model, we use 0, 2, 4, 6, and 9

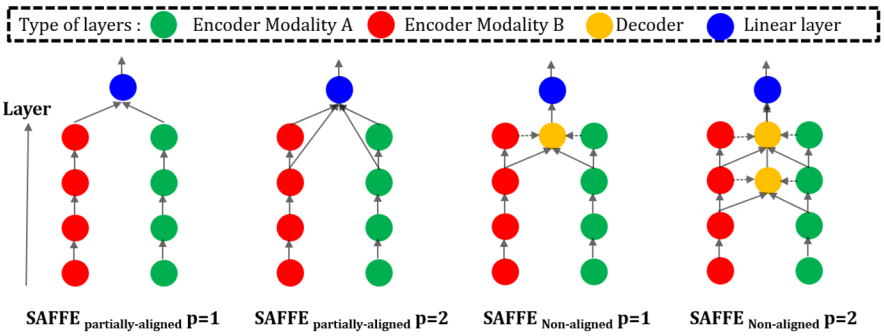


Fig. 5 SAFFE-derived experimental model for analysis. $\text{SAFFE}_{\text{Partially-aligned}}$ (left): Fusion with single encoder (P=1). $\text{SAFFE}_{\text{Partially-aligned}}$(middle, left): Fusion with consecutive encoder layers (P=2). $\text{SAFFE}_{\text{Non-aligned}}$(middle, right): Fusion with single encoder (P=1). $\text{SAFFE}_{\text{Non-aligned}}$ (right): Fusion with consecutive encoder layers (P=2). The solid arrows indicate the outputs from the encoder, while the dashed arrows signify the corresponding weights (K, Q, V)

encoder layers with consecutive layers, depicted as $0+1, 2+3, 4+5, 6+7$, and $9+10$ in Figures 6a through 8c. Below, we analyze each of these figures.

Figure 6a and Figure 6b presents the outcomes for the $\text{SAFFE}_{\text{Partially-aligned}}$ and $\text{SAFFE}_{\text{Non-aligned}}$ models, respectively, demonstrating the performance metrics (i.e., the mAP metric) associated with the datasets under the SAFFE configuration, as shown in Table 2. To demonstrate the importance and successful alignment of the semantic space of both input modalities of our trainable FAM, the figures additionally show the mAP results when the output from the frozen encoder model (which incorporates all encoder layers) is used to directly compute the cosine similarity between modalities in the absence of the SAFFE methodology. These results are represented in the figures as SAFFE_FREE. As we can see, Figure 6a and 6b shows that without the SAFFE methodology, accuracy declines significantly. In contrast, $\text{SAFFE}_{\text{Partially-aligned}}$ and $\text{SAFFE}_{\text{Non-aligned}}$ models achieve higher mAP scores across all datasets.

More specifically, in Figure 6a, the $\text{SAFFE}_{\text{Partially-aligned}}$ model shows that the individual encoder layer ($P = 1$) $h^{11} = 11$ (final hidden state) obtains the highest mAP value, whereas the concatenation of encoder layers ($P = 2$) $h^{10} = 10$ and $h^{11} = 11$ (last two hidden states) yields the optimal mAP. Figure 6b illustrates the outcomes of the $\text{SAFFE}_{\text{Non-aligned}}$ model, wherein the individual encoder layer ($P = 1$) fusion mechanism, characterized by a layer $h^{10} = 10$, yields the highest mAP value. Furthermore, the concatenation of the two layers ($P = 2$) within the mechanism, specifically layers $h^9 = 9$ and $h^{10} = 10$ (represented as 9+10), results in the maximum mAP value. Among these models, the two-layer ($P = 2$) concatenation demonstrates the highest mAP value. As a result, it suggests that this synergy

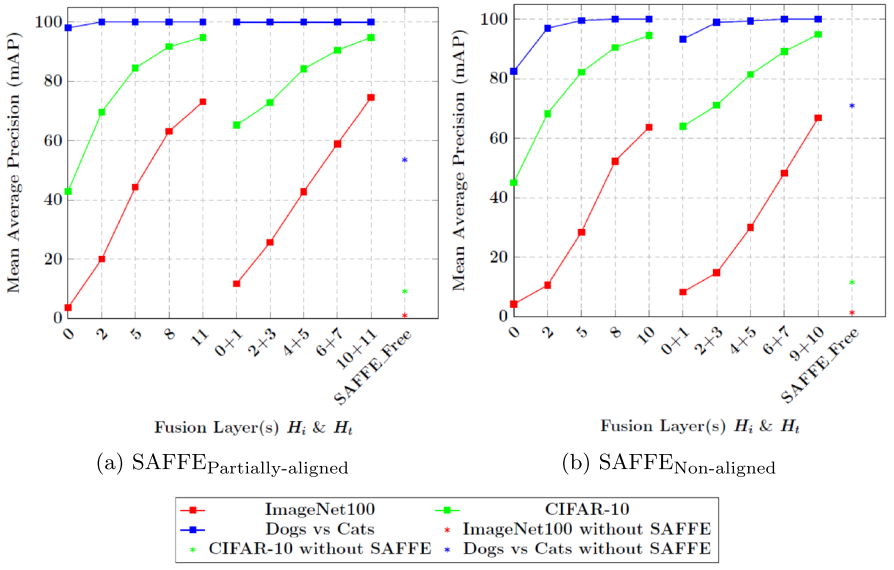


Fig. 6 (a) Partially-aligned and non-aligned (b) fusion of encoders, wherein both text and image encoders facilitate either one encoder layer ($P = 1$), e.g., 11 in X-axis, or two consecutive encoder layers ($P = 2$), e.g., $10 + 11$ in X-axis. The mAP for the SAFFE_Free, absent the SAFFE-derived model

supports performance optimization, indicating potential avenues for further exploration of encoder layer combinations.

Takeways: The final encoder layers effectively gather and condense feature representations from the input modalities, enabling a robust representation for the semantic alignment and fusion of FAM between modalities. Furthermore, this indicates that the collaborative interaction between these layers enhances overall performance. Furthermore, our experiments indicate that the optimal number of fusion layers in the SAFFE methodology is $P = 2$. Additional experiments for $P > 2$ (not included in this paper) do not improve accuracy of this downstream task for the evaluated datasets with the selected frozen encoders. This finding simplifies the model significantly while also drastically lowering computational costs and enhancing performance.

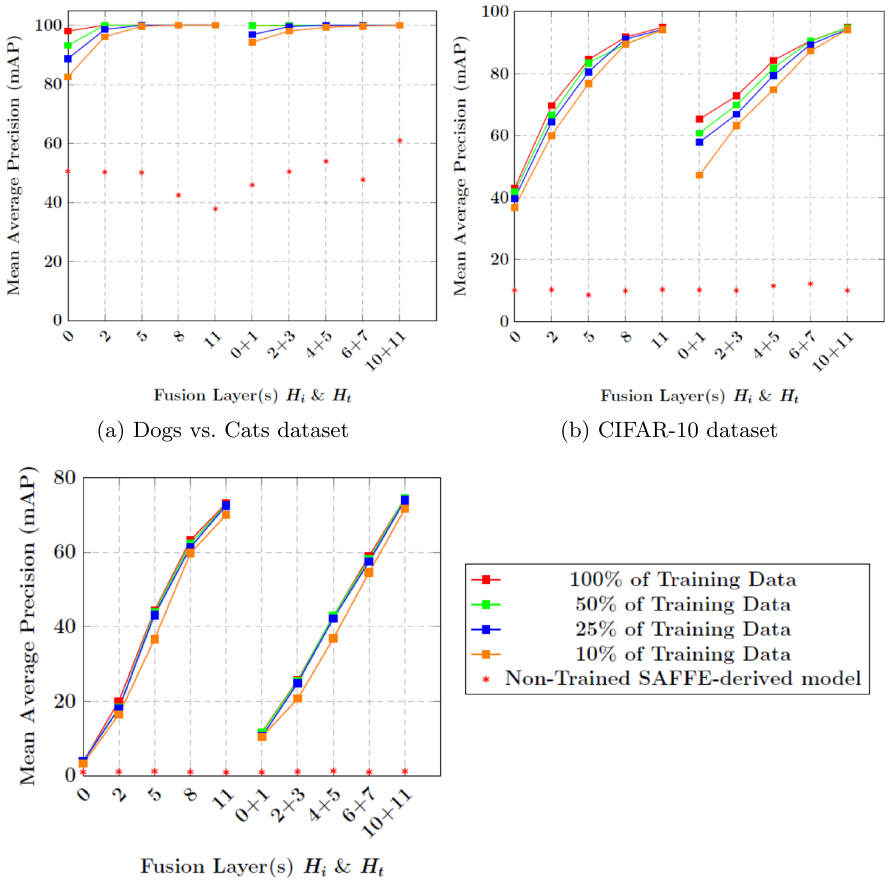
Analyzing the results across the different datasets, we observe that in the case of Dogs vs. Cats dataset, Figure 6a shows that the $\text{SAFFE}_{\text{Partially-aligned}}$ mode attains a peak accuracy (100% mAP) at encoder layer $h^2 = 2$ for $P = 1$ and at layers $h^2 = 2$ and $h^3 = 3$ for $P = 2$. Likewise, Figure 6b demonstrates that the $\text{SAFFE}_{\text{Non-aligned}}$ mode achieves a maximum accuracy (100% mAP) at layer $h^5 = 5$ for $P = 1$ and at layers $h^6 = 6$ and $h^7 = 7$ for $P = 2$. These results demonstrate that image retrieval tasks involving simple datasets such as Dogs vs. Cats dataset do not need the execution of all encoder layers (12 layers) from the pre-trained model; rather, the initial few encoder layers provide sufficient contextual information for this purpose.

Takeways: *This finding suggests that optimizing the use of encoder layers can lead to more efficient model performance, allowing for faster training times and reduced computational resources while maintaining high accuracy in retrieval tasks.*

4.2.2 Semantic-alignment fusion with scarce training data

This study aims to evaluate the performance accuracy of a SAFFE-derived model in relation to the volume of training data used as shown in Table 1. Figure 7 shows the outcomes for the $\text{SAFFE}_{\text{Partially-aligned}}$ models in comparison with the $\text{SAFFE}_{\text{Non-aligned}}$, as illustrated in Figure 8, using the SAFFE configurations listed in Table 2 with the datasets Dogs vs. Cats, CIFAR-10, and ImageNet100. In addition, these figures present the optimal mAP values along with the accuracy metrics for the non-trained SAFFE-derived model with random weights. To illustrate the rapid adaptability of the SAFFE-derived model, we performed training using 100%, as well as 50%, 25%, and 10% of the datasets. We report mAP across all scenarios, adhering to established models.

As we can see, for the three datasets, Figure 7a, 7b, and 7c (output from the $\text{SAFFE}_{\text{Partially-aligned}}$ model) and Figure 8a, 8b, and 8c (output from the $\text{SAFFE}_{\text{Partially-aligned}}$ models) demonstrate that the accuracy of the model without training the SAFFE-derived model is considerably low, especially as the complexity of the dataset increases. However, as expected, when the SAFFE-derived model is trained, much higher mAP values are achieved across all datasets. All the graphical representations indicate that training on the complete dataset



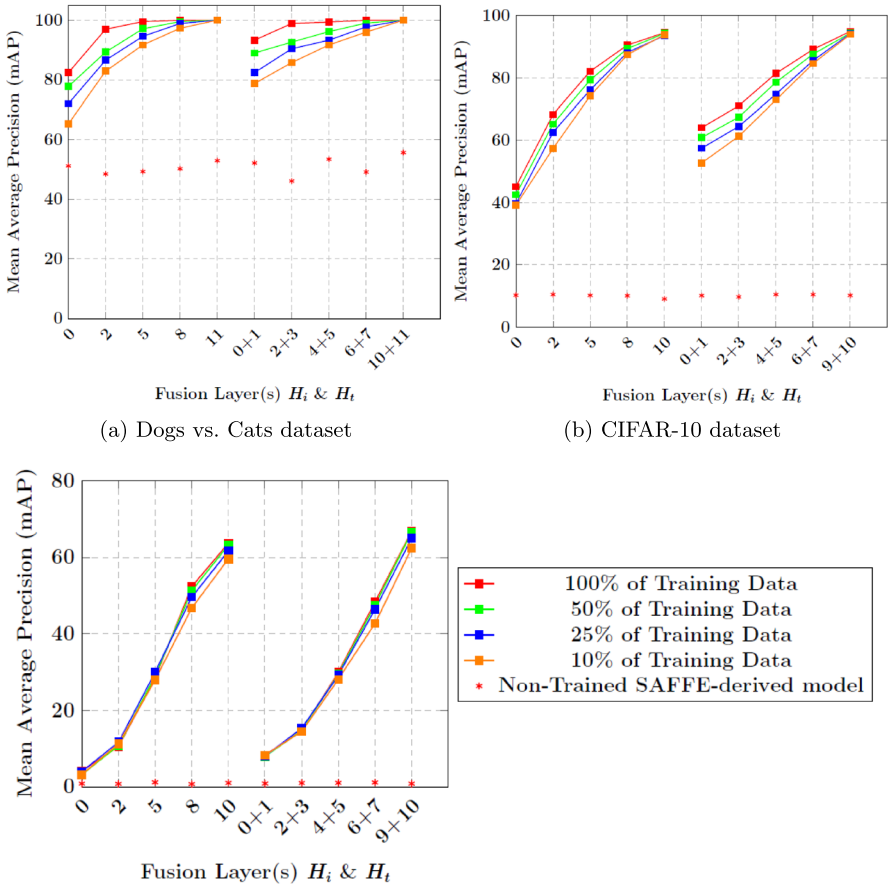
(a) Dogs vs. Cats dataset

(b) CIFAR-10 dataset

(c) ImageNet100 dataset

Fig. 7 Training with the Dogs vs. Cats (a), CIFAR-10 (b), ImageNet100 (c) dataset at proportions of 100%, 50%, 25%, and 10% of the data, utilizing partially-aligned encoders, where both text and image encoders provide one ($P = 1$) or two encoder layers ($P = 2$) for fusion, results in the highest mAP performance alongside the non-trained SAFFE-derived model accuracy at the inference phase

(100%) yields the mAP values across all datasets with both $\text{SAFFE}_{\text{Partially-aligned}}$ and $\text{SAFFE}_{\text{Non-aligned}}$ models. In the case of the SAFFE-derived model utilizing a single encoder input ($P = 1$), encoder 11 achieves the highest mAP values for $\text{SAFFE}_{\text{Partially-aligned}}$ models, while encoder 10 attains the maximum mAP values for $\text{SAFFE}_{\text{Non-aligned}}$ models. When the SAFFE-derived model is implemented with a consecutive dual encoder input ($P = 2$), encoders 10 and 11 exhibit the highest mAP values for $\text{SAFFE}_{\text{Partially-aligned}}$ model, whereas encoders 9 and 10 secure the maximum mAP values for $\text{SAFFE}_{\text{Non-aligned}}$ models. Notably, when the proportion of training data is diminished, the corresponding highest mAP values experience only a marginal reduction. In scenarios involving $\text{SAFFE}_{\text{Partially-aligned}}$ and $\text{SAFFE}_{\text{Non-aligned}}$ models with $P = 1$ and $P = 2$, it is observed that utilizing



(a) Dogs vs. Cats dataset

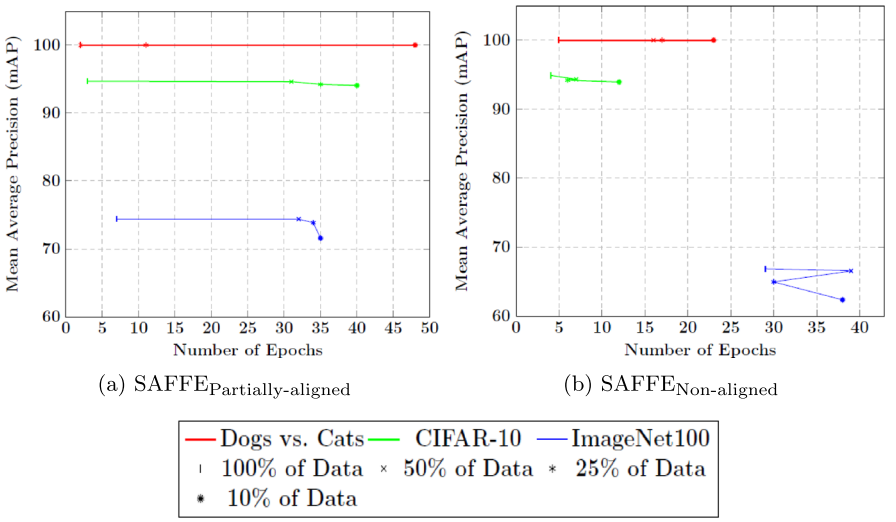
(b) CIFAR-10 dataset

(c) ImageNet100 dataset

Fig. 8 Training with the Dogs vs. Cats (a), CIFAR-10 (b), ImageNet-100 (c) dataset at proportions of 100%, 50%, 25%, and 10% of the data, utilizing non-aligned encoders, where both text and image encoders provide one or two encoder layers for fusion, results in the highest mAP performance alongside the non-trained SAFFE-derived model accuracy at the inference phase

merely 10% of the training dataset can approximate a performance level similar to that achieved with the entire dataset (100%). This observation underscores that the model not only demonstrates efficiency but also effectiveness in harnessing limited data to optimize performance.

Next, in Figure 9a and 9b, we analyze the number of epochs required for training convergence in each case, considering the highest-performing $\text{SAFFE}_{\text{Partially-aligned}}$ model (10 + 11) and $\text{SAFFE}_{\text{Non-aligned}}$ model (9 + 10). Both Figure 9a and 9b show that 50% of the training dataset attains mAP values that are remarkably close to those of the complete dataset training values following an increase in the number of training epochs to fit the models more effectively. The increment of training epochs (at 50% of data) for the $\text{SAFFE}_{\text{Partially-aligned}}$ model consisted of 25 epochs



(a) $\text{SAFFE}_{\text{Partially-aligned}}$ (b) $\text{SAFFE}_{\text{Non-aligned}}$

Fig. 9 (a) Partially-aligned and (b) non-aligned encoder fusion model utilizes the concatenation of the hidden states 10 with 11 (10+11) and 9 with 10 (9+10), respectively, resulting in the highest mAP attained by the agents corresponding to the respective echo number

for ImageNet100 and 28 epochs for CIFAR-10, with no requirement for the Dogs vs. Cats dataset. Conversely, the supplementary training epochs for the more challenging $\text{SAFFE}_{\text{Non-aligned}}$ model comprised 10 epochs for ImageNet100, 3 epochs for CIFAR-10, and 11 epochs for the Dogs vs. Cats dataset. Additionally, 10% of the dataset also approaches slightly below the 50% training dataset threshold with an increased number of training epochs. The dataset comprising two classes, Dogs vs. Cats, attained the maximum mAP value of 100% utilizing merely 10% of the dataset after an increase in the number of training epochs.

Takeways: *Our performance analysis of SAFFE-derived models trained on different subsets of the full training dataset shows that they can achieve competitive mAP metrics, even in the worst-case scenario (training with only 10%). In general, the smaller the training dataset, the more epochs are required for training. However, this increase is not significant, with fewer than 30 epochs in the worst case (10%) for the most complex and compute-intensive training (ImageNet100 dataset). This finding implies that the model's design is not only robust but also rapidly adaptable, enabling efficient generalization even when faced with minimal information. This highlights the potential for deploying SAFFE-derived models in scenarios where data scarcity is a challenge, such as in specialized medical applications or emerging fields with limited datasets.*

4.2.3 Encoder layers contribution for semantic-alignment fusion

This study aims to evaluate the performance accuracy of a SAFFE-derived model in relation to the number of encoders contributing to the FAM unit ($P = 1, 2, 3, 4$). Table 3 shows the results for the $\text{SAFFE}_{\text{Non-aligned}}$ model utilizing the SAFFE

settings outlined in Table 2 with the datasets Dogs vs. Cats, CIFAR-10, and ImageNet100. The table provides the highest mAP value along with the corresponding epoch number to the related P values.

According to the findings presented in the table, the configuration comprising two encoder layers ($P = 2$) with the FAM unit demonstrates the most effective attainment of a favorable mPA value with a less number of epochs across all datasets. Although configurations with $P > 2$ reach optimal mAP values, they necessitate greater computational power for model training. Thus, the most appropriate number of encoders is two ($P = 2$) from each modality for the FAM unit. These results highlight the importance of balancing model complexity and computational efficiency, as augmenting the number of encoder layers beyond two may yield marginal performance improvements but incur considerable expenses in terms of training duration and resource allocation.

Takeways: The results demonstrate a substantial relationship between the number of encoder layers and the efficacy of the model. In accordance with this, the optimal configuration of the FAM unit, which includes two encoder layers from each modality ($p = 2$), improves accuracy across various datasets.

4.2.4 SAFFE-derived models versus frozen SOTA multimodal models

As explained in Section 2, SOTA multimodal models possess zero-shot learning capabilities due to their large number of trainable parameters, Transformer-based architecture, and the extensive datasets used for training. This allows them to be effectively deployed for inference even on unseen or partially seen datasets. Table 5 presents a comparative analysis of zero-shot best performance across various SOTA bimodal models on the CIFAR-10 and CIFAR-100 datasets. As shown in the table, these models can achieve high mAP values, particularly on the simpler CIFAR-10 dataset, without explicit and costly fine-tuning of their numerous trainable parameters (see the second column of the table). However, for a specific downstream task such as image retrieval, where the more complex CIFAR-100 dataset is required for training, fine-tuning these SOTA models becomes highly computationally expensive.

To enable more efficient fine-tuning of existing multimodal models for specific datasets required in a target downstream task, end-users can leverage our SAFFE methodology. More specifically, we first select the frozen encoder components of interest and then train only the FAM unit using the targeted dataset.

Table 3 The SAFFE framework integrates novel dual modalities with pre-existing frozen models. We performed experiments using varying counts ($p = 1, 2, 3, 4$) of encoder layers and achieved the highest mAP value at the corresponding epoch number

Dataset	$P = 1$	$P = 2$	$P = 3$	$P = 4$
ImageNet100	63.68%/37	66.84%/29	66.90%/39	68.12%/39
CIFAR-10	94.46%/14	94.88%/4	94.98%/9	95.02%/12
Dogs vs. Cats	100%/3	100%/5	100%/8	100%/20

Table 4 The configurations of SAFFE-derived models along with their corresponding frozen models. *EL*=Number of encoder layers and *ED*=Embedding dimension

Model	Text Encoder	Image Encoder	EL	ED
$SAFFE_{Non-aligned_{01}}$	sentence-transformers/all-mpnet-base-v2	laion/CLIP-ViT-B-32-laion2B-s34B-b79K	12	768
$SAFFE_{Non-aligned\_large}$	sentence-transformers/all-roberta-large-v1	openai/clip-vit-large-patch14	24	1024
$SAFFE_{Non-aligned\_MAE}$	sentence-transformers/all-roberta-large-v1	facebook/vit-mae-base	12	768

Table 5 mAP performance of SOTA bimodal models and SAFFE-derived models

Model	# Trainable Parameters	CIFAR-10	CIFAR-100
OpenCLIP [34]	1B	93.5	76.2
CN-CLIP [70]	958 M	96.0	79.7
ALIGN [10]	820 M	94.9	76.8
CLIP [9]	400 M	94.9	77.0
Wukong [68]	307 M	95.4	77.1
ZLaP [40]	151 M	93.6	73.3
$SAFFE_{Non-aligned\_MAE}$	264 M	81.32	44.6
$SAFFE_{Partially-aligned}$	67 M	94.72	75.38
$SAFFE_{Non-aligned\_01}$	67 M	96.6	77.46
$SAFFE_{Non-aligned\_large}$	351 M	**97.39**	**80.94**

This new training approach offers two key advantages. First, by training only the FAM unit, the computational cost is significantly lower compared to fine-tuning the entire stack of encoder and decoder layers in the SOTA model. Second, our SAFFE methodology provides full flexibility in composing the fine-tuned multimodal model by allowing the selection of per-modality encoder layers that are most efficient for the target downstream task. To demonstrate both benefits, we compare the SOTA models with two SAFFE-derived models: $SAFFE_{partially-aligned}$, configured as shown in Table 2. $SAFFE_{Non-aligned}$, configured as shown in Table 4.

As shown in Table 5, in terms of trainable parameters, our SAFFE-derived models require significantly fewer trainable parameters than the SOTA models while achieving competitive mAP values on both the CIFAR-10 and CIFAR-100 datasets. The Vision-Language Pre-training model proposed by Wukong [68] comprises 100 million pairs of Chinese image text gathered from the Internet. The ZLaP [40] represents a contemporary model characterized by a minimal number of trainable parameters. This results in substantially lower computational costs for fine-tuning. More importantly, the flexibility of our SAFFE methodology allows for the composition of a higher-performing SAFFE-derived model—specifically, the non-aligned variant—by strategically selecting different types of frozen encoders. Notably, the $SAFFE_{Non-aligned\_01}$ model, which incorporates

a more effective selection of frozen encoders, namely, CLIP and SBERT, in an off-the-shelf manner, achieves significantly higher mAP than the partially-aligned model. This model achieves the highest accuracy when compared to the SOTA model while utilizing the fewest number of trainable parameters (67 M). Furthermore, the $\text{SAFFE}_{\text{Non-aligned\_large}}$ model achieves the highest mAP value attained by all SOTA and SAFFE models. This model encompasses 351 M parameters; however, it does not surpass the model with the highest accuracy, namely, the CN-Clip model(958 M). The encoder models that have been selected are not explicitly trained on the target datasets; rather, the chosen pre-trained models extract features from the input data and generate a sophisticated vector space through their extensive pre-training processes. The FAM unit has the capability to seamlessly integrate the encoder layers and produce a new modality-invariant vector representation for the input data. This subsequent fusion of encoder layers with FAM can enhance the accuracy of the downstream task.

The $\text{SAFFE}_{\text{Non-aligned\_MAE}}$ experiment aimed to demonstrate the FAM unit's fusion ability and efficiency with various unimodal encoders. This model integrates the fusion of the MAE-ViT [69] image encoder with the SBERT text encoder within its architecture. The MAE [69] (Masked Autoencoder) is a straightforward autoencoding approach that reconstructs the original signal from its partial observations. We employed the MAE-ViT encoder as illustrated in Table 4, utilizing the pre-trained MAE-ViT encoder that has been trained on the ImageNet1K dataset. Furthermore, we conducted experiments utilizing the CIFAR-10 and CIFAR-100 datasets, as displayed in result Table 5. The SAFFE model demonstrates exceptional accuracy when utilizing the MAE-ViT encoder, even in the absence of prior training on those datasets. In this experimental setup, both encoders contribute two encoder layers ($P = 2$) to the FAM unit, with the datasets being exclusively trained using the FAM unit. This illustrates the capability of democratic pre-trained encoders and models for improved efficacy. It underscores the adaptability and efficiency of the SAFFE model. The SAFFE model showcases robustness across different types of image data, suggesting its adaptability in various real-world applications.

Takeways: *The SAFFE methodology enables the efficient fusion of frozen encoders and flexible composition of multimodal models, enhancing mAP performance for a targeted downstream task without the costly fine-tuning of all layers in existing SOTA multimodal models on the datasets of interest. This approach optimizes resource utilization by allowing the model to leverage existing knowledge without extensive training, facilitating faster and more efficient deployment in real-world applications.*

4.2.5 SAFFE-derived models versus SOTA fusion-based models

As explained in Section 2.3, certain SOTA bimodal models, such as BLIP and UniT, utilize efficient mid-fusion mechanism while MAGMA implements a late fusion mechanism. Since our FAM unit implements a bottleneck mid-fusion technique, we aim to assess its effectiveness by comparing two $\text{SAFFE}_{\text{Non-aligned}}$

Table 6 mAP performance of SOTA mid-fusion-based bimodal models and $SAFFE_{Non-aligned}$ models in image retrieval with COCO dataset. VG: Visual Genome, CC: Conceptual Captions, and SBU: SBU Captions datasets

Model/#Total Parameters	Pre-train Dataset	# Trainable Parameters	Text backbone	Image backbone	Trainable Encoders/Decoders	Training Time/Hardware	mAP
Flamingo [18]/80B	M3W dataset (185 M images)	70B	BERT	NFNet	86/86	15 days/TPUv4	65.9
SigLIP [37]/1.8B	WebLI/10B	1.8B	Transfor mer-based	ViT-B/16	24/0	5 days/32 × TPU-v4 chips	65.1
ALIGN [10]/411 M	1.8B	411 M	$BERT_L$	Efficient Net-B7	24 encoder layers & 813 layers	Several days/16 × 1024 Cloud TPUv3 cores	59.9
BLIP [20]/361 M	COCO + VG + CC + SBU (14 M images)	361 M	$BERT_{base}$	VIT−B/16	13/1	8 days/8× A100-80 G GPU [71]	63.1
ALBEF [72] /209.5M	COCO + VG + CC + SBU (14 M images)	209.5M	$BERT_{base}$	VIT−B/16	24/1	8 days/8× NVIDIA A100 GPU	60.7
UniT [36]/201 M	COCO (1.5 M images)	201 M	$BERT_{base}$	DETR [73]	24/12	3 days/8×V100	40.13
MAGMA [21]/6180 M	CC-3 M (3.3 M)	13 M	GPT-J [31]	CLIP	Projection Layer	1.75 days/16×A100 GPU	52.1
$SAFFE_{Non-aligned\_01}$ /899 M	COCO (1.5 M images)	FAM (240 M)	SBERT	VIT−B/16	FAM (0/2)	2.5 days/RTX-3060 GPU	56.35
$SAFFE_{Non-aligned\_02}$ /726 M	COCO (1.5 M images)	FAM (67 M)	SBERT	VIT−B/32	FAM (0/2)	2.5 days/RTX-3060 GPU	**67.2**

models with these SOTA fusion models for image retrieval tasks using the COCO dataset (details in Table 6).

As we can see in the table, the model $\text{SAFFE}_{\text{Non-aligned\_01}}$ employs a VIT−B/16 vision encoder, which is characterized by a sequence length of 197 with 16 pixel size patches, whereas another model ($\text{SAFFE}_{\text{Non-aligned\_02}}$) utilizes a VIT−B/32 vision encoder with a sequence length of 49 with 32 pixel size patches. In both SAFFE-derived models, the FAM unit has the $P = 2$ configuration using the last encoder layers 9 and 10 of the two (image and text) frozen encoders. In terms of computational cost, the VIT-B/32 ($\text{SAFFE}_{\text{Non-aligned\_02}}$) architecture necessitates considerably less computational power due to its reduced sequence length, consequently achieving the highest mAP value.

Compared to SAFFE, in spite of their advanced fusion strategies (Section 2.3), the seven evaluated SOTA models generally require a significantly higher number of trainable encoders and decoders. In contrast, SAFFE-derived models adopt a lightweight trainable architecture (FAM) that includes only two decoders coupled with a linear layer. This streamlined configuration reduces the need for pairwise attention mechanisms and substantially decreases the total number of trainable parameters. For instance, the best-performing SAFFE model includes just 67 million parameters ($P = 2$) and achieves the highest mAP (67.2). Models such as BLIP [20], Flamingo [18], and UniT [36] depend on full pairwise attention across all input modalities to reach their peak mAP. However, this approach demands significantly more computational resources, as illustrated in Table 6. Likewise, MAGMA [21] employs a massive 6 billion parameter decoder-only GPT-J model with 4096-dimensional hidden states, further increasing complexity and compute load. Similarly, SigLIP [37] and ALIGN [10] models feature notable requirements as well, as including the need for extensive computational power and massive datasets for end-to-end training, which can take several days to complete.

Takeaways: *SAFFE-derived models—featuring fewer trainable parameters and efficient cross-modal fusion via the FAM unit—significantly reduce the computational cost of training compared to the seven evaluated SOTA models, which frequently necessitate expensive, comprehensive full pairwise attention mechanisms. SAFFE employs later layer bottleneck mid-fusion techniques to circumvent the necessity for extensive training. Consequently, SAFFE models' training can converge much faster using commodity GPU hardware, making them practical for a wider range of users.*

4.2.6 SAFFE-derived models and fast concept binding

The fast concept binding of visual and linguistic components means the capacity to acquire proficiency in a novel language task after being prompted with merely a limited number of instances within few-shot learning. This study specifically examines how accurately fast concept binding can be achieved using SAFFE-derived models and compares their performance to the state-of-the-art Few-Shot Learning (FSL) model [52].

Following the experimental methodology of the FSL work, we perform image classification with the miniImageNet dataset [38] and adopt the same terminology:

- **Number of ways:** Number of object classes in the classification (e.g., dog vs. cat).
- **Number of inner shots:** Number of unique instances per category (i.e., the number of images in the dog class).

Image class labels have been substituted with meaningless terms such as ("Blorbin," "Crundle," etc.) that correspond to actual items ("cat," "dog," etc.). We initiated this experiment since nonsensical terms carry no (or minimal) intrinsic meaning, to evaluate the extent of difficulty introduced by associating visual categories with these meaningless words with limited inner shots with few training epochs. Text encoders remain "blind" to modalities beyond textual representation, thereby restricting our ability to convey visual context to them. For our SAFFE-derived models, we maintain the encoder in a frozen state and trained the FAM unit using both 2-way and 5-way configurations with 1, 3, and 5 inner shots over 1 or 2 epochs.

Table 7 presents the experimental results for the 2-way and 5-way classifications. The Meta-Learning [41] framework yields outcomes that surpass those achieved by the FSL frozen [52] model in both the 2 and 5 ways of the experimental design under 1 and 5 inner shots. By leveraging a SAFFE-derived model with an improved combination of frozen text and image encoders—compared to the rigid architecture of the meta-learning—we achieve a higher accuracy in both settings. Additionally, training for just two epochs significantly improves accuracy.

We also conducted two more complex experiments: one using 20 distinct superclasses from the CIFAR-100 dataset and another with the ImageNet100 dataset, where all 100 class names were replaced with nonsensical terms. After training FAM with both image and nonsensical word modalities, the model derived from SAFFE attained a high accuracy of 81.89 on the CIFAR-100 dataset and 60.3 on the ImageNet dataset. The CIFAR-100 superclass experiment further demonstrates

Table 7 Performance of fast concept binding with miniImageNet in 2-way and 5-way tasks. *Number of epochs for training

Number of ways		**2**			**5**		
Inner shots	**Image/Text backbone**	**1**	**3**	**5**	**1**	**3**	**5**
FSL $\text{frozen}_{\text{test-blind}}$ [52] 1*	NF-ResNet-50/Transformer Architecture [22]	48.5	46.7	45.3	18.6	19.9	19.8
Meta-Learning [41] 1*	CLIP-ViT/B-32/GPT-2	58.7	-	65.8	25.1	-	29.6
$\text{SAFFE}_{\text{Non-aligned}}$ 1*	CLIP-ViT/B-16/SBERT	80.0	78.0	77.0	30.4	54.0	46.4
$\text{SAFFE}_{\text{Non-aligned}}$ 2*	CLIP-ViT/B-16/SBERT	95.0	90.0	96.0	30.0	80.0	74.8

strong generalization across multiple image categories, providing empirical evidence for our proposed methodology's effectiveness in integrating new linguistic data.

From the previous experiments, we observe that a SAFFE-derived model can learn new image name associations when presented with images alongside their corresponding descriptions. The model's ability to use newly introduced words improves with additional examples from the same category, and increasing training epochs further enhances accuracy. The SAFFE-derived model leverages a pre-trained language model to extract linguistic features, enabling rapid learning and seamless interaction between visual and textual modalities—without modifying the weights of the frozen encoders. FAM training bridges the gap between these modalities, allowing the trained FAM to generate class names for corresponding images using the new linguistic model.

Takeways: The FAM unit rapidly binds various words in languages that incorporate visual elements by utilizing pre-trained frozen text and image encoders. This capability can facilitate the acquisition of a range of new tasks, as example represented as a sequence of several interleaved image and text embeddings. This development can understand and interpret information about images with various aspects and provide associated information. As well, it opens up new avenues for applications in fields such as automated content creation, accessibility tools for the visually impaired, and enhanced human–computer interaction [52].

5 Conclusions and future work

In this work, we present SAFFE, a novel methodology for the flexible and scalable composition of multimodal models, specifically tailored to evolving end-user downstream tasks. In contrast with existing multimodal models and state-of-the-art fusion techniques, SAFFE-derived models eliminate the need for expensive end-to-end training or full fine-tuning to achieve high accuracy on target datasets. SAFFE leverages per-modality off-the-shelf frozen encoders—readily available from major AI providers—by selectively integrating only those components necessary for the downstream task. This targeted selection avoids over-parameterization and significantly reduces the model's memory footprint. Since these pre-trained frozen encoders are often trained independently and not within a unified multimodal context, their output embeddings may be semantically misaligned. To resolve this, we propose the FusionAlign Module (FAM)—a lightweight, bottleneck mid-fusion unit trained solely on the target end-user dataset. FAM aligns the semantic spaces across modalities, enabling effective multimodal integration without updating the parameters of the frozen encoders.

As a proof of concept, we demonstrate the effectiveness of the SAFFE methodology in a bimodal setting involving image and text modalities, applied to image retrieval and language-based downstream tasks. Through extensive experiments and ablation studies, we evaluate a range of SAFFE fusion strategies that combine various types of frozen encoders—both partially aligned and fully non-aligned—across datasets of varying complexity. Our results show that SAFFE can flexibly and efficiently compose high-accuracy

bimodal models, achieving improved prediction performance compared to state-of-the-art methods, while significantly reducing computational costs.

Future research will explore the integration of a third modality—specifically audio—while optimizing the FAM unit design to enhance performance and cost efficiency in SAFFE-based tri-modal models incorporating text, image, and audio inputs. These enhanced SAFFE-derived models will enable new downstream tasks, such as image segmentation and Visual Question Answering particularly by leveraging the synergy between text and image modalities.

Acknowledgments This work has been funded by MICIU/AEI/10.13039/501100011033 and by "European Union NextGenerationEU/PRTR" under the grants CNS2023-144241 and RYC2021-031966-I.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Akbari H, Yuan L, Qian R, Chuang W-H, Chang S-F, Cui Y, Gong B (2021) Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems 34:24206–24221
2. Gu Z, Lang B, Yue T, Huang L (2017) Learning joint multimodal representation based on multi-fusion deep neural networks. In: Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24, pp. 276–285 Springer
3. Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C (2021) Attention bottlenecks for multimodal fusion. Advances in neural information processing systems 34:14200–14213
4. Girdhar R, Singh M, Ravi N, Van Der Maaten L, Joulin A, Misra I (2022) Omnivore: A single model for many visual modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16102–16112
5. Guzhov A, Raue F, Hees J, Dengel A (2022) Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 976–980 IEEE
6. Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I (2023) Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180–15190
7. Zhang Y, Gong K, Zhang K, Li H, Qiao Y, Ouyang W, Yue X (2023) Meta-transformer: A unified framework for multimodal learning. arXiv preprint arXiv:2307.10802
8. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755 Springer
9. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 PMLR
10. Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, Le QV, Sung Y, Li Z, Duerig T (2021) Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision arxiv: 2102.05918

11. Liu J, Mao Y, Huang Z, Ye Y (2023) A bottleneck network with light attention for multimodal clustering. Knowledge-Based Systems 280:111037
12. Shvetsova N, Chen B, Rouditchenko A, Thomas S, Kingsbury B, Feris RS, Harwath D, Glass J, Kuehne H (2022) Everything at once-multi-modal fusion transformer for video retrieval. In: Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition, pp. 20020–20029
13. Li Y, Jiang S, Hu B, Wang L, Zhong W, Luo W, Ma L, Zhang M (2024) Uni-moe: Scaling unified multimodal llms with mixture of experts. arXiv preprint arXiv:2405.11273
14. Piergiovanni A, Noble I, Kim D, Ryoo MS, Gomes V, Angelova A (2024) Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26804–26814
15. Xie S, Sun C, Huang J, Tu Z, Murphy K (2018) Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 305–321
16. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803
17. Chen Q, Wang W, Huang K, De S, Coenen F (2021) Multi-modal generative adversarial networks for traffic event detection in smart cities. Expert Systems with Applications 177:114939
18. Alayrac J-B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M et al (2022) Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems 35:23716–23736
19. Koh JY, Salakhutdinov R, Fried D (2023) Grounding language models to images for multimodal inputs and outputs. In: International Conference on Machine Learning, pp. 17283–17300 PMLR
20. Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900 PMLR
21. Merullo J, Castricato L, Eickhoff C, Pavlick E (2023) Linearly mapping from image to text space. In: The Eleventh International Conference on Learning Representations
22. Vaswani A (2017) Attention is all you need. Advances in Neural Information Processing Systems
23. Boulahia SY, Amamra A, Madi MR, Daikh S (2021) Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications 32(6):121
24. Yang Z, Wang J, Tang Y, Chen K, Zhao H, Torr PH (2022) Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18155–18165
25. Zheng G, Zhou X, Li X, Qi Z, Shan Y, Li X (2023) Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22490–22499
26. Wu T, Li X, Qi Z, Hu D, Wang X, Shan Y, Li X (2024) Spherediffusion: Spherical geometry-aware distortion resilient diffusion model. Proceedings of the AAAI Conference on Artificial Intelligence 38:6126–6134
27. Guo Y, Gao L, Song J, Wang P, Sebe N, Shen HT, Li X (2021) Relation regularized scene graph generation. IEEE Transactions on Cybernetics 52(7):5961–5972
28. Li X, Zheng G, Yu Y, Ji N, Li X (2024) Relationship-incremental scene graph generation by a divide-and-conquer pipeline with feature adapter. IEEE Transactions on Image Processing
29. Morency L-P, Baltrušaitis T (2017) Multimodal machine learning: integrating language, vision and speech. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 3–5
30. Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M et al (2022) Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems 35:25278–25294
31. Wang B, Komatsuzaki A (2021) GPT-J-6B: A 6 billion parameter autoregressive language model
32. Morency L-P, Baltrušaitis T (2017) Multimodal machine learning: integrating language, vision and speech. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 3–5
33. Wang W, Tran D, Feiszli M (2020) What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12695–12705

34. Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, Schuhmann C, Schmidt L, Jitsev J (2023) Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2818–2829
35. Chen S, He X, Guo L, Zhu X, Wang W, Tang J, Liu J (2023) Valor: Vision-audio-language omni-perception pretraining model and dataset. arXiv preprint arXiv:2304.08345
36. Hu R, Singh A (2021) Unit: Multimodal multitask learning with a unified transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1439–1449
37. Zhai X, Mustafa B, Kolesnikov A, Beyer L (2023) Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11975–11986
38. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 Ieee
39. Krizhevsky A (2009) Learning multiple layers of features from tiny images. Technical report
40. Kalantidis Y, Tolias G et al (2024) Label propagation for zero-shot classification with vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23209–23218
41. Najdenkoska I, Zhen X, Worring M (2022) Meta-learning makes a better multimodal few-shot learner. In: Sixth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems
42. Devlin J (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
43. Reimers N (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084
44. Liu Y (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692**364**
45. Dosovitskiy A (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
46. Gong Y, Chung Y-A, Glass J (2021) Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778
47. Yao M, Tao D, Gao R, Qi P (2024) Anomaly detection for mec enabled hierarchical industrial iot with transformer enhanced variational auto encoder. IEEE Transactions on Industrial Informatics
48. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846
49. Liao J, Shi Y, Gong M, Shou L, Qu H, Zeng M (2021) Improving zero-shot neural machine translation on language-specific encoders-decoders. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 IEEE
50. Cho Y, Yu H, Kang S-J (2023) Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. IEEE Transactions on Multimedia
51. Ding H, Liu C, Wang S, Jiang X (2021) Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16321–16330
52. Tsimpoukelli M, Menick JL, Cabi S, Eslami S, Vinyals O, Hill F (2021) Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems 34:200–212
53. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2022) Transformers in vision: A survey. ACM computing surveys (CSUR) 54(10s):1–41
54. Pang Z, Xie Z, Man Y, Wang Y-X (2023) Frozen transformers in language models are effective visual encoder layers. arXiv preprint arXiv:2310.12973
55. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971
56. Liu Z (2024) Improving the inference efficiency of transformer models in machine translation tasks by low-rank decomposition methods. In: 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS), pp. 930–936 IEEE
57. Qin B, Li J, Tang S, Zhuang Y (2025) Dba: Efficient transformer with dynamic bilinear low-rank attention. IEEE Transactions on Neural Networks and Learning Systems
58. Chen Y, Shang J, Zhang Z, Sheng J, Liu T, Wang S, Sun Y, Wu H, Wang H (2024) Mixture of hidden-dimensions transformer. arXiv preprint arXiv:2412.05644

59. Cukierski W (2013) Dogs vs. Cats. https://kaggle.com/competitions/dogs-vs-cats. Kaggle
60. CIFAR 10. (2021) https://huggingface.co/datasets/uoft-cs/cifar10. Hugging Face
61. Shekhar A (2021) ImageNet100. https://www.kaggle.com/datasets/ambityga/imagenet100. Kaggle
62. CIFAR 100. (2021) https://huggingface.co/datasets/uoft-cs/cifar100. Hugging Face
63. COCO. (2021) https://huggingface.co/datasets/detection-datasets/coco. huggingface
64. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**
65. Wu B, Xu C, Dai X, Wan A, Zhang P, Yan Z, Tomizuka M, Gonzalez J, Keutzer K, Vajda P (2020) Visual Transformers: Token-based Image Representation and Processing for Computer Vision
66. Wolf T (2019) Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771
67. Kingma DP (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
68. Gu J, Meng X, Lu G, Hou L, Minzhe N, Liang X, Yao L, Huang R, Zhang W, Jiang X et al (2022) Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. Advances in Neural Information Processing Systems 35:26418–26431
69. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009
70. Yang A, Pan J, Lin J, Men R, Zhang Y, Zhou J, Zhou C (2022) Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335
71. Jian Y, Liu T, Tao Y, Zhang C, Vosoughi S, Yang H (2023) Expedited training of visual conditioned language generation via redundancy reduction. arXiv preprint arXiv:2310.03291
72. Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH (2021) Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems 34:9694–9705
73. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 Springer

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Maithri Kulasekara[1,3] · Juan F. Inglés-Romero[4] · Baldomero Imbernón[2] · José L. Abellán[1]

✉ Maithri Kulasekara
dm.kulasekara@um.es

Juan F. Inglés-Romero
jf.ingles@libelium.com

Baldomero Imbernón
bimbernon@ucam.edu

José L. Abellán
jlabellan@um.es

[1] Dept. of Computer Engineering and Technology, Universidad de Murcia, Murcia 30100, Spain

[2] Dept. of Computer Science, Universidad Católica de Murcia, Murcia 30107, Spain

[3] Computer Engineering Dept., General Sir John Kotelawala Defence, Rathmalana, Sri Lanka

[4] Emerging Tech Dept., Libelium LAB, Urban Center, Murcia 30006, Spain